TALN 2009, Senlis, 24–26 juin 2009

Nouveau paradigme d'évaluation des systèmes de dialogue homme-machine

Marianne Laurent[1] Ghislain Putois[1] Philippe Bretier [1] Thierry Moudenc[1]
(1) Orange Labs, Lannion, 2 avenue Pierre Marzin, 22307 Lannion Cedex

Résumé. L'évaluation des systèmes de dialogue homme-machine est un problème difficile et pour lequel ni les objectifs ni les solutions proposées ne font aujourd'hui l'unanimité. Les approches ergonomiques traditionnelles soumettent le système de dialogue au regard critique de l'utilisateur et tente d'en capter l'expression, mais l'absence d'un cadre objectivable des usages de ces utilisateurs empêche une comparaison entre systèmes différents, ou entre évolutions d'un même système. Nous proposons d'inverser cette vision et de mesurer le comportement de l'utilisateur au regard du système de dialogue. Aussi, au lieu d'évaluer l'adéquation du système à ses utilisateurs, nous mesurons l'adéquation des utilisateurs au système. Ce changement de paradigme permet un changement de référentiel qui n'est plus les usages des utilisateurs mais le cadre du système. Puisque le système est complètement défini, ce paradigme permet des approches quantitatives et donc des évaluations comparatives de systèmes.

Abstract. Evaluation of a human-machine dialogue system is a difficult problem for which neither the objectives nor the proposed solutions gather a unanimous support. Traditional approaches in the ergonomics field evaluate the system by describing how it fits the user in the user referential of practices. However, the user referential is even more complicated to formalise, and one cannot ground a common use context to enable the comparison of two systems, even if they are merely an evolution of the same service. We propose to shift the point of view on the evaluation problem : instead of evaluating the system in interaction with the user in the user's referential, we will now measure the user's adequacy to the system in the system referential. This is our Copernician revolution : for the evaluation purpose, our system is no longer user-centric, because the user referential is not properly objectifiable, while the system referential is completely known by design.

Mots-clés : Évaluation, Dialogue.

Keywords: Evaluation, Dialogue.

1 Introduction

La question de l'évaluation est aujourd'hui un enjeu majeur pour la recherche en dialogue vocal. Pour preuve, nombreuses sont les manifestations aujourd'hui dédiées à ce sujet (Jokinen *et al.*, 2007) (McTear *et al.*, 2008). L'étude critique réalisée par (Paek, 2007; Paek, 2001) a identifié deux perspectives différentes face à cette problématique. D'un côté, la recherche académique s'attache à identifier quelle est la meilleure démarche d'évaluation pour comparer précisément des systèmes de dialogue entre eux. De l'autre, l'industrie, qui est confrontée à des objectifs opérationnels de déploiement en situation réelle (contrainte de coût et nécessité de qualité), réfléchit à des outils pour l'aider à concevoir de meilleurs systèmes de dialogue vocal.

De façon générale, avec les outils dont elle dispose, l'industrie n'a qu'une vue partielle de la qualité de ses services. Certes elle sait collecter une grande variété de mesures, de manière à la fois quantitative sur des sélections d'indicateurs, et qualitative pour estimer la satisfaction utilisateur, mais elle bute sur la définition de processus qui puissent répondre de manière fiable à ses besoins. Ce manque de métriques de qualité complique la tâche des concepteurs, mais nuit également à l'essor des systèmes de dialogue vocal par frilosité des entreprises susceptibles de les déployer. Aussi, en marge de la recherche d'un paradigme universel d'évaluation pour la comparaison entre systèmes, nous présentons ici une méthode pragmatique proposant aux concepteurs des clés pour évaluer leur système face aux attentes opérationnelles.

Dans les sections suivantes, nous présentons notre contexte industriel de développement de système de dialogue, puis nous introduisons une nouvelle approche au problème de l'évaluation, et présentons enfin des méthodes d'évaluation associées à ce changement d'approche, et qui viennent soutenir l'industrie tout au long du cycle de vie de ses systèmes.

2 Développement industriel d'un système de dialogue vocal

Les systèmes de dialogue vocaux sont des objets complexes. Aussi, leur évaluation, qui, selon l'objectif, peut intervenir à tout moment du cycle de développement, requiert la prise en compte de différents points de vue (en particulier logiciel, fonctionnel, pragmatique, sociologique). Posons tout d'abord quelques définitions associées à ces points de vue. Nous désignons par *système de dialogue* la plateforme matérielle et les composantes logicielles associées : moteurs de reconnaissance et synthèse vocales, gestion des lignes téléphoniques, composants de compréhension et de génération en langue naturelle, moteur de dialogue. La *logique de dialogue* désigne la logique suivie par le système de dialogue lors de ses interactions avec l'utilisateur. Elle est dédiée à une tâche et suit une logique métier définie pour l'accomplir. Enfin, l'*application* englobe l'ensemble formé par le système de dialogue et la logique de dialogue.

Chez Orange, la conception industrielle d'une application de dialogue s'organise en quatre étapes. Tout d'abord, la phase de réalisation correspond à la création d'une première version de l'application. Puis celle-ci est ensuite progressivement améliorée par les phases d'expérimentation et pilote au travers une séries d'itérations. Enfin la phase d'exploitation marque son déploiement. À chaque étape on effectue une analyse des interactions menées entre l'application et ses utilisateurs de sorte, notamment, à identifier les écarts entre les comportements utilisateurs attendus et ceux observés, tant en matière de type de comportement (typologie de réponses et de questions) que d'occurrence (fréquence, délais) ou de forme (choix lexicaux, forme syntaxique). Cette analyse nourrit ainsi les itérations qui correspondent à des retours

de l'application en conception durant lesquels on modifie son comportement en fonction des expériences observées dans les traces de dialogue.

La phase de réalisation s'articule autour des étapes de conception (réflexions, spécifications, modélisation), de développement et de test de l'application au sein même de l'équipe de développement. Celle-ci met en place tous les aspects d'une version initiale de l'application : design, linguistique, ergonomie et technique (notamment lié à l'intégration de l'application dans son environnement informatique). L'évaluation à cette étape repose sur des tests effectués par l'équipe de développement et constitue une évaluation du bon fonctionnement technique.

Lors de la phase d'expérimentation, l'application est testée par un nombre restreint d'utilisateurs puis ouvert à un faible flux de trafic réel. On collecte alors la matière nécessaire à l'analyse des comportements utilisateur face à l'application. Cette analyse permet d'alimenter les boucles d'itérations qui vont se succéder pour améliorer l'application jusqu'à l'atteinte d'un niveau de performance jugé satisfaisant. La collecte et l'analyse du corpus d'interaction permettent ainsi de capturer des usages et des réactions des utilisateurs et donc d'adapter les différents paramètres du design, tels que les modèles de reconnaissance vocale, la syntaxe et la terminologie des prompts ou la temporisation du dialogue. L'évaluation à cette étape vise donc l'amélioration de l'application et notamment l'alignement des prévisions et des observations.

Lors de la phase pilote, les tests sont effectués dans les conditions réelles d'exploitation. D'abord, le panel d'utilisateurs peut couvrir une zone géographique (souvent région ou département), une zone téléphonique, une zone de plateau de téléconseillers ou un panel de clients cibles, mais elle couvre toujours une partie du flux réel. Ensuite, l'évaluation est réalisée à partir de l'architecture technique complète de l'application. Cette étape correspond donc à une phase de beta-testing où les problèmes majeurs sont décelés et comblés avant la mise en production qui ouvre l'application à l'ensemble des usagers. L'évaluation à cette étape à pour but de valider globalement l'application.

Enfin, la phase d'exploitation correspond à la mise à disposition totale de l'application avec maintien opérationnel. L'évaluation se focalise alors sur la supervision. Cette évaluation permet notamment d'analyser la façon dont l'application est utilisée en situation réelle par les utilisateurs, d'analyser l'usage et les retours d'expérience utilisateurs et leurs évolutions dans le temps.

3 Changement de paradigme

Comme pour tout processus d'évaluation, évaluer une application de dialogue nécessite un référentiel d'étude. Les tentatives traditionnelles cherchent à appréhender les usages des utilisateurs pour mesurer l'adéquation de l'application à ces usages (Norros & Savioja, 2007). Ces méthodes d'évaluation recensées par (Grislin & Kolski, 1996) et (Paek, 2001) s'articulent en deux étapes. D'abord on collecte un corpus de travail par le biais d'études qualitatives telles que des tests terrains, des interviews, des questionnaires de satisfaction ou des simulations en Magicien d'Oz (Caelen *et al.*, 1997). Puis les ergonomes interprètent des usages et des attentes utilisateurs à partir de ce corpus pour former le référentiel d'étude.

Les pratiques actuelles sont focalisées sur les besoins de la recherche, à savoir un outil précis pour le benchmark de solutions. Par exemple, un paradigme répandu pour l'évaluation comparative d'applications de dialogue homme-machine est PARADISE de (Walker *et al.*, 1997). Il

consiste à réaliser des mesures quantitatives sur les traces d'interaction d'une application afin d'approximer la satisfaction utilisateur par une combinatoire de métriques quantitatives. Ces méthodes cherchent à bâtir leur référentiel d'évaluation sur l'utilisateur humain, ce qui rend la tâche complexe et difficilement objectivable. La constitution de ce référentiel devient donc artisanale, spécifique à une application donnée, et en conséquence, la tâche doit être réitérée à chaque nouvelle campagne d'évaluation. Pour autant, étudier une application nécessite de prendre en compte ses utilisateurs, puisque c'est en interagissant avec eux que l'application prend tout son sens et devient objet propice à l'évaluation.

Nous avons bien conscience qu'une évaluation se fait toujours par rapport à un cadre, mais nous souhaitons ici rappeler que dans le domaine des systèmes de dialogue, chaque interaction entre l'utilisateur et l'application présente le caractère unique d'une performance. Chaque nouveau dialogue est co-construit différemment selon les capacités propres à la fois à l'application et à l'utilisateur, ce dernier sachant s'adapter très vite, et changer ses pratiques. Pour pouvoir réaliser une évaluation comparative entre deux applications ou entre deux évolutions d'une même application, nous avons impérativement besoin que ce cadre d'évaluation soit stable et applicable aux deux. Or les pratiques utilisateurs ne sont pas assez stables pour constituer ce cadre, justement à cause des grandes facultés d'adaptation humaines.

Dans l'industrie, les exigences de commensurabilité au regard de l'évaluation sont moins grandes. L'industrie cherche avant tout à mesurer l'adéquation d'une application face aux attentes opérationnelles (contraintes de coût et nécessité de qualité). En se plaçant dans cette logique industrielle, il nous semble donc pertinent d'inverser notre approche sur la relation entre application et utilisateur pour l'évaluation. Au lieu de tenter de mesurer à quel point l'application est proche de l'utilisateur et de ses pratiques, envisageons plutôt de déterminer dans quelle mesure l'utilisateur est proche de l'application et des pratiques qu'elle propose. Le cadre d'évaluation à prendre en compte est alors beaucoup plus stable, car l'application s'adapte beaucoup moins vite que l'utilisateur, et seulement dans la mesure où on la fait évoluer ou on la dote de capacités d'apprentissage (toujours assez limitées). Elle est donc entièrement maîtrisable, car l'application est conçue pour répondre à un ensemble de besoins fonctionnels.

Parmi ces besoins, il y a la nécessité à traiter les comportements utilisateurs (des mots aux enchaînements d'actes de langage). Le niveau technologique actuel ne permettant pas de traiter tous les usages de la langue, le design nécessite de circonscrire les comportements à considérer, c'est-à-dire de prédire les comportements utilisateurs. De plus, la connaissance issue des études centrées utilisateur sert à la prédiction des comportements utilisateurs, et *in fine* à la définition des besoins fonctionnels. Ces besoins, comme dans tout processus industriel, doivent également pouvoir se traduire par un ensemble de métriques pour vérifier qu'ils sont adressés. De plus, ils sont les invariants principaux qui définissent la raison d'être d'une application, et donc ils restent majoritairement valables tout au long du cycle de vie de l'application, ce qui permet d'utiliser le même cadre d'évaluation pour comparer deux évolutions d'une même application.

4 Les mesures industrielles

Pour refonder notre évaluation, plaçons-nous maintenant dans le référentiel de l'application qui a été défini dans la section précédente. Dans ce nouveau référentiel, il est possible d'extraire des caractéristiques récurrentes de l'application et des mesures objectivables associées, qui constitueront la matière première pour l'évaluation quantitative des applications. Ainsi, pour évaluer

la pertinence d'une application de dialogue, nous choisissons des indicateurs qui mesurent l'accomplissement de la tâche et la manière dont le dialogue est mené avec l'utilisateur. Des corpus de mesure sont ainsi constitués à partir d'une sélection d'indicateurs. Une multitude d'indicateurs peuvent ainsi être sollicités, parmi lesquels : les performances de chaque composant logiciel, le nombre d'appels au service, le nombre moyen de tours de dialogue pour accomplir une tâche, la durée de chaque tour de dialogue, le nombre et l'instant des raccrochés en cours de dialogue et les taux d'incompréhension de la reconnaissance vocale et des composants de compréhension du langage naturel.

Pour passer de la mesure à l'évaluation, il faut définir un cadre d'évaluation. On sélectionne d'abord un ensemble d'indicateurs qui apparaissent *a priori* représentatifs du bon fonctionnement de l'application, puis on associe des interprétations aux ensembles de valeurs possibles de ces indicateurs. Le cadre ainsi construit est indépendant du système initial, il définit des seuils sur les indicateurs. Ce cadre d'évaluation peut alors être appliqué à un autre système de dialogue ou à une version alternative de l'application.

Rappelons que des processus d'évaluation interviennent à plusieurs endroits distincts du cycle de vie d'une application. Ceci implique autant de redéfinitions d'un cadre d'évaluation propre aux métiers concernés (design, marketing, exploitation, etc.), en fonction de leurs objectifs respectifs et des moyens de mesure à leur disposition. Ainsi, lors de la phase de réalisation, les designers vérifient le comportement nominal de l'application par rapport aux pratiques qu'ils ont préalablement définies dans le cahier des charges. L'application n'étant bien souvent testée ni sur l'architecture cible, ni par des utilisateurs réels, seuls quelques indicateurs clés sont utilisés. Ensuite, on mobilise davantage d'indicateurs pour les phases d'expérimentation et pilote afin d'évaluer l'écart entre les usages réels et les prédictions d'usages tels qu'implémentés dans l'application. Enfin, en phase d'exploitation, seuls quelques indicateurs clés sont utilisés pour la supervision. L'application étant censée bien fonctionner, on vérifie alors principalement que les utilisateurs continuent d'interagir avec l'application conformément aux possibilités d'interaction.

Nous voyons donc comment les processus de conception de l'application et de conception de son évaluation peuvent être étroitement entremêlés dans le cadre d'un développement industriel. Ceci est d'autant plus vrai que les spécifications de l'application se focalisent sur des aspects locaux de l'application. A titre d'exemple, notre application automatique de renseignement par téléphone propose une mise en communication avec l'interlocuteur recherché moyennant une surtaxe. Une évaluation locale sur les taux de réponses positives à la dernière question de l'application a fait opter les designers pour le prompt système «Pour être mis en relation, dites oui.» qui avait un meilleur score que son prédécesseur «Souhaitez-vous être mis en relation ?».

5 Conclusion

Consciente des économies que les systèmes de dialogue vocal peuvent générer, et de l'impact de la satisfaction client en termes d'image, l'industrie des services cherche donc à rationaliser l'évaluation de ses applications pour pouvoir maîtriser leur exploitation et leurs usages.

Les utilisateurs ont de fortes capacités d'adaptation aux contraintes imposées par les applications. Nous suggérons donc de ne pas chercher à évaluer les applications de dialogue uniquement avec des méthodes centrées utilisateur, mais au contraire d'exploiter la construction des systèmes autour d'un ensemble d'usages prédéfinis, et de chercher alors à évaluer si les utili-

sateurs s'accommodent des usages proposés. La conception des évaluations est alors plus aisée pour chaque phase de décision du cycle de vie d'un système, puisqu'elle se réfère directement à des objectifs définis dans le cahier des charges, eux-mêmes quantifiables par des indicateurs de performance du système évalué.

La méthode proposée répond donc à un besoin industriel. Elle repose sur des optimisations locales dans le but de réduire l'écart entre les interactions observées entre l'application et ses utilisateurs et celles attendues. Ainsi, si elle intègre la notion de progrès dans la comparaison de plusieurs versions d'une même application, elle n'adresse cependant pas la question de recherche d'un optimum global de qualité, notion que nous ne savons pas définir aujourd'hui.

Remerciements

Nous souhaitons remercier les équipes Dialogue et Synthèse Vocale, les équipes multidisciplinaires et opérationnelles d'Orange pour leur contribution à ces réflexions.

Références

CAELEN J., ZEILIGER J., BESSAC M., SIROUX J. & PÉRENNOU G. (1997). Les corpus pour l'évaluation du dialogue homme-machine. In *Actes des 1ères JST FRANCIL 1997, Journées Scientifiques et Techniques du Réseau Francophone de l'Ingénierie de la Langue de l'AUPELF-URE, Avignon, 15/04/97-16/04/97*, p. 215–222 : -.

GRISLIN M. & KOLSKI C. (1996). Evaluation des interfaces homme-machine lors du développement des systèmes interactifs = human-machine interface evaluation during the development of interactives systems. *TSI. Technique et science informatiques ISSN 0752-4072 CODEN TTSIDJ*, **15**(3), 265–296.

JOKINEN K., MCTEAR M. & LARSON J. (2007). Dialogue on dialogues – multidisciplinary evaluation of advanced speech-based interactive systems : A report on the interspeech 2006 satellite event. *AI Magazine*, **28**(2).

MCTEAR M., JOKINEN K. & LARSON J. (2008). Special issue on evaluating new methods and models for advanced speech-based interactive systems. *Speech Communication*, **50**(8-9).

NORROS L. & SAVIOJA P. (2007). Vers une théorie et une méthode d'évaluation de l'utilisabilité des systèmes complexes homme-technologie. *@ctivités*, **4**(2).

PAEK T. (2001). Empirical methods for evaluating dialog systems. In *Proceedings of the workshop on Evaluation for Language and Dialogue Systems*, p. 1–8, Morristown, NJ, USA : Association for Computational Linguistics.

PAEK T. (2007). Toward evaluation that leads to best practices : Reconciling dialog evaluation in research and industry. In *Proceedings of the Workshop on Bridging the Gap : Academic and Industrial Research in Dialog Technologies*, p. 40–47, Rochester, NY : Association for Computational Linguistics.

WALKER M. A., LITMAN D. J., KAMM C. A. & ABELLA A. (1997). Paradise : a framework for evaluating spoken dialogue agents. In *Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics*, p. 271–280, Morristown, NJ, USA : Association for Computational Linguistics.