Constitution d'un corpus de dialogue oral pour l'évaluation automatique de la compréhension hors et en contexte du dialogue

H Maynard [1], K. McTait[2] , D. Mostefa[2],
L. Devillers [1], S. Rosset[1], P.Paroubek[1] , C. Bousquet [3], K. Choukri[1], J. Goulian[4], J-Y Antoine [5],
F. Béchet [6], O. Bontron[7], L. Charnay [8], L. Romary [9], M. Vergnes[10], N. Vigouroux [3]

media@elda.fr

[1]LIMSI, [2]ELDA/ELRA, [3]IRIT, [4]CLIPS, [5]VALORIA, [6]LIA, [7]TELIP, [8]FRANCE-TELECOM R&D, [9]LORIA, [10]VECSYS

ABSTRACT

This paper presents and reports on the progress of the EVALDA/MEDIA project, focusing on the recording protocol of the reference dialogue corpus. The aim of this project is to define and test an evaluation methodology that assess and diagnose the context-sensitive understanding capability of spoken language dialogue systems. Systems from both academic organizations (CLIPS, IRIT, LIA, LIMSI, LORIA, VALORIA) and industrial sites (FRANCE TELECOM R&D, TELIP) will be evaluated. ELDA is the coordinator of the Technolangue/EVALDA multi-campaign evaluation project, a national initiative sponsored by the French government, of which MEDIA is a sub-campaign. MEDIA began in January 2003. VECSYS provides the recording platform for the project.

1. INTRODUCTION

L'objectif du projet MEDIA de l'action TECHNOLANGUE est de définir et de tester une méthodologie d'évaluation de la compréhension hors et en contexte des systèmes de dialogue. Nous proposons de mettre en place un paradigme d'évaluation fondé sur la définition et l'utilisation de batteries de tests issues de corpus réels et sur une représentation sémantique et des métriques communes. Ce paradigme devrait permettre de diagnostiquer les capacités de compréhension hors contexte et en contexte des systèmes de dialogue. Ce paradigme sera utilisé dans le cadre d'une campagne d'évaluation qui réunira les systèmes des différents sites, sur une même tâche de demandes de renseignements.

Actuellement, il n'existe pas de méthodologie standard, ni même de pratique communément admise dans la communauté scientifique, pour évaluer et comparer des systèmes de dialogue. La nature dynamique et interactive du dialogue rend difficile la constitution d'un jeu de données de test afin d'offrir un référentiel d'évaluation commun à plusieurs tâches. Par contre, des projets de taille conséquente ont tenté de jeter les bases d'une méthodologie d'évaluation pour les systèmes de dialogue oral, en commençant par le projet francophone AUF- Arc B2 [1], l'évaluation par DEFI [2], les projets européens EAGLES [3], DISC [4], SUNDIAL [5] ainsi que les projets ATIS [6] et maintenant COMMUNICATOR [7] aux USA.

Le paradigme PEACE (Paradigme d'Evaluation Automatique de la Compréhension hors et En contexte dialogique) [8,9] sur lequel est fondé le projet MEDIA permet une évaluation automatique, comparative et diagnostique pour la compréhension hors et en contexte dialogique. Il est fondé sur la constitution de batteries de tests reproductibles issues de dialogues réels. Ce paradigme suit le même courant d'idée que les évaluations DQR [10] et DEFI [2] basées sur des batteries de tests. L'environnement d'évaluation repose sur l'idée, que dans le cadre de systèmes portant sur des tâches de renseignements liées à une base de données, il est possible de mettre en place une représentation sémantique commune, vers laquelle chaque système est capable de convertir sa propre représentation. De plus, le paradigme permet une évaluation en contexte du dialogue. Le contexte est simulé de façon artificielle, par une paraphrase, le but étant de tester l'interprétation d'un énoncé U dans le contexte D^n (pour reprendre les notations de l'approche DQR). Enfin, alors que les grands programmes d'évaluation étaient centrés sur l'évaluation des performances (mesures globales), cette campagne devrait permettre non seulement une évaluation des performances mais aussi un diagnostic des modélisations utilisées.

L'objectif du projet MEDIA est donc de donner à la communauté scientifique francophone les moyens d'évaluer comparativement les approches de la compréhension, en lui offrant la possibilité de partager des corpus et en définissant des représentations et des métriques génériques communes. La première étape du projet MEDIA a été consacrée à la définition et à la constitution d'un corpus commun de dialogues en français, dédié à la tâche retenue dans MEDIA (serveur d'information touristique). Après une présentation du projet MEDIA, l'article présente la méthodologie utilisée pour la collecte du corpus (définition de la

tâche, description de la plate-forme d'enregistrement, protocole), ainsi que les premières observations sur ce corpus.

2. LE PROJET MEDIA

2.1 Organisation de la campagne

L'organisation d'une campagne d'évaluation de la compréhension hors et en contexte des systèmes de dialogue a pour but principal de promouvoir une dynamique de l'évaluation au sein de la communauté. L'objectif de ce projet est de mettre en place un paradigme générique d'évaluation pour la compréhension hors et en contexte dialogique permettant une évaluation automatique, comparative et diagnostique de systèmes.

Une campagne d'évaluation doit garantir la pérennité des ressources constituées pour la campagne ainsi que les produits dérivés de celle-ci. Pour garantir l'impartialité de la campagne, l'évaluation doit être menée par un partenaire qui ne participe pas à la campagne. ELDA prend en charge cet aspect et enregistre le corpus nécessaire pour ce projet, il se charge de produire ou de faire produire les outils nécessaires à l'évaluation. De plus, il prend en charge l'organisation de la campagne et l'évaluation des résultats. La société VECSYS a mis en place la plate-forme d'enregistrement du corpus (matériel et outil WOZ). Le LIMSI en tant que promoteur du projet a le rôle de coordinateur scientifique.

Les participants aux évaluations sont aussi bien des partenaires académiques (CLIPS, IRIT, LIA, LIMSI, LORIA, VALORIA) qu'industriels (FRANCE TELECOM R&D, TELIP).

2.2 Paradigme d'évaluation

Afin de permettre une évaluation diagnostique, le paradigme d'évaluation s'appuie sur une représentation générique commune.

Représentation sémantique générique commune

Il s'agit de mettre en place une représentation du sens des énoncés utilisateurs permettant d'établir une relation d'équivalence dans l'ensemble des requêtes possibles. Une réflexion est menée pour définir cette représentation commune en dehors de tout domaine, mais dans le contexte de systèmes de renseignements liés à une base de données. Le formalisme de représentation choisi doit être consensuel et permettre l'annotation de grands corpus. Il est basé sur une structure attributs-valeurs qui permet la représentation de structures complexes. Ce formalisme permet aussi bien de coder l'acte de dialogue que le contenu propositionnel d'un énoncé. Il est convenu que chaque participant prend en charge la conversion depuis sa propre représentation interne vers la représentation commune.

Unités de référence

Une unité de référence pour l'évaluation de la compréhension hors contexte, comprend la transcription exacte des énoncés utilisateurs et la représentation sémantique de référence. Une unité de références pour l'évaluation de la compréhension en contexte de dialogue, comprend le contexte sous forme d'une paraphrase [8], la transcription exacte de l'énoncé utilisateur et la représentation sémantique résultant de l'interprétation de l'énoncé compte tenu du contexte. La paraphrase peut être obtenue soit à partir des annotations en contexte du corpus, soit par la concaténation des phrases usagers et des réponses du système.

L'ensemble des unités de référence sera divisé en trois parties : un corpus d'adaptation à la tâche (10k requêtes), un corpus de développement (2k requêtes) distribué aux partenaires et un corpus de test caché (3k requêtes) pour l'évaluation. Chaque énoncé utilisateur, transcrit suivant les normes de transcriptions des énoncés oraux, est annoté suivant la représentation sémantique commune hors contexte et en contexte.

Mesures d'évaluation communes

L'objectif est de définir des mesures communes permettant d'effectuer des évaluations diagnostiques des systèmes. Il doit être possible également de pondérer l'importance des erreurs selon les types établis ci-dessous.

Définition et typologie des phénomènes et fonctions dialogiques

Le paradigme doit offrir une analyse qualitative et diagnostique automatique des performances du module de compréhension hors contexte et en contexte. On pourra par exemple s'intéresser à des difficultés particulières de l'oral qui existent hors contexte: hésitation, répétition etc. Une liste des fonctionnalités de la compréhension en contexte à tester, s'inspirant de la documentation réalisée dans le cadre de la campagne par DEFI [2], sera définie par le consortium (ellipses, anaphores, relâchement de contraintes etc.).

3. CONSTITUTION DU CORPUS

3.1 Définition de la tâche et du domaine

Dans le cadre d'une campagne d'évaluation de dialogues homme-machine, nous avons décidé de restreindre la tâche aux applications de demandes de renseignements accédant à des bases de données, telles que serveur touristique, serveur d'horaires de train, d'avion etc. La définition de la représentation sémantique est générique. Elle est ensuite adaptée à la tâche et à la base de données. L'idéal était de travailler sur une application reliée à une base de données réelle, par exemple à partir d'un accès au WEB sur un site d'agence de voyage ou d'office de tourisme. La tâche commune choisie pour l'évaluation est celle

d'informations touristiques concernant la réservation d'hôtels à partir de sites web.

3.2 Collecte du corpus

Nous avons besoin de corpus de dialogue communs pour adapter les différents systèmes de compréhension en contexte dialogique et pour créer les batteries de tests servant à l'évaluation. Pour que l'évaluation ne soit pas biaisée, nous avons décidé d'enregistrer un nouveau corpus en utilisant une simulation de serveur vocal d'informations touristiques par magicien d'Oz. Ainsi chaque locuteur croit dialoguer avec une machine alors que le dialogue est en réalité pris en charge par un humain (un « compère ») qui simule les réponses d'un serveur d'informations touristiques. Ceci nous permet d'obtenir un corpus de dialogues variés, grâce notamment aux comportements du compère.

Dans cette campagne, il est prévu de ne travailler que sur les transcriptions exactes des dialogues comprenant les transcriptions des intervenants (utilisateurs et système). Cependant, il nous semble important de disposer également du signal audio numérisé de bonne qualité correspondant aux dialogues, afin de pouvoir élargir la campagne au traitement d'entrées issues d'un système de reconnaissance de la parole.

La taille envisagée du corpus est d'environ 15000 requêtes utilisateurs. Pour cela 1250 dialogues provenant de 250 locuteurs seront enregistrés, chaque locuteur effectuant 5 scénarii différents. Le corpus final contiendra de l'ordre de 70 heures d'enregistrement.

Plateforme d'enregistrement

La méthode choisie de collection du corpus est celle du « magicien d'Oz » (Wizard of Oz : WoZ). Celle-ci consiste à simuler un dialogue homme-machine en langage naturel. La simulation provient du fait que la machine est remplacée par une personne qui répond à une demande de l'utilisateur en imitant le fonctionnement automatique d'un serveur vocal. Pour ce faire, cet opérateur ou compère utilise un outil graphique développé par VECSYS, qui l'aide à la génération des réponses qu'il doit communiquer à l'appelant. Les phrases de génération sont obtenues en complétant un modèle avec les informations provenant d'un site Web d'informations touristiques et les données de l'appelant. La Figure 1 illustre le fonctionnement de l'outil d'enregistrement.

Le signal est enregistré directement au format numérique. Les dialogues sont ensuite transcrits orthographiquement, segmentés en actions dialogiques et annotés sémantiquement.

Protocole d'enregistrement

Les utilisateurs participant aux enregistrements se basent sur des scenarii de réservation de chambres d'hôtel qui ont été générés à partir de scénarii de base de façon à avoir une diversité dans les dialogues. Afin d'obtenir des requêtes utilisateurs présentées de

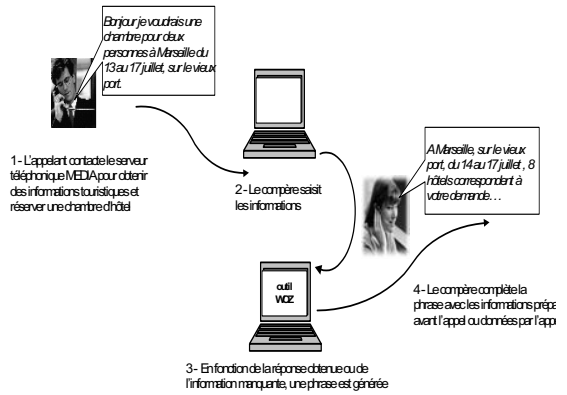

Figure 1 : fonctionnement de l'outil WoZ

manière la plus naturelle possible, ces scénarii sont communiqués aux utilisateurs par téléphone afin de réduire les paraphrases langagières des textes du scénarii.

Plusieurs points d'entrée dans le dialogue sont possibles : choix d'une ville, choix d'un itinéraire, choix d'un événement, prix, période. Huit catégories de scénarii ont été définies pour avoir des niveaux de complexité différente. Un exemple d'un scénario complexe est indiqué en Figure 2. comprenant la réservation de plusieurs hôtels à plusieurs endroits selon un itinéraire particulier.

DATE:	du 20/02 au 25/02
LIEU:	du 20/02 au 21/02 (1 nuit) à Lille, du 21/02 au 23/02 (2 nuits) à Paris et du 23/02 au 25/02 (2 nuits) à Marseille
NB-CHAMBRES	2 couples, un avec 1 enfant
NB-ADULTES	4
NB-ENFANTS	2
PRIX:	bon standing (maxi 200€)
DIVERS:	Mercure, animaux, parking

Figure 2 : Exemple de scénario

En plus de la variété des scénarii fournis aux locuteurs, nous avons défini des consignes que le compère utilise pour répondre aux requêtes de l'utilisateur. Le premier type de consigne concerne les erreurs de reconnaissance vocale ou de compréhension. Ainsi le compère produira une réponse basée sur une erreur de compréhension de la requête utilisateur. Le deuxième type concerne les confirmations implicites ou explicites que formule le compère. Enfin un dernier type de consigne concerne le niveau de coopération du compère. Si la consigne est d'être coopératif, le compère répondra et donnera toutes les informations à l'utilisateur. Inversement, le compère peut être non coopératif et ne pas répondre ou répondre partiellement aux requêtes de l'utilisateur. En plus des consignes compère, le locuteur reçoit des consignes sur le

nombre et le type de critères qu'il peut négocier avec le serveur.

3.3 Etat du corpus

Ce corpus est en cours d'enregistrement. Actuellement, plus des 4/5èmes du corpus a été enregistré, c'est-à-dire environ 1000 dialogues. La Table 1 indique des mesures moyennes sur un échantillon de 200 dialogues transcrits. On remarque notamment que les énoncés locuteurs restent relativement courts (6 mots).

Durée moyenne	3.4 minutes
Enoncés locuteur par dialogue	15 énoncés
Enoncés compère par dialogue	18 énoncés
Longueur d'un énoncé locuteur	6 mots
Longueur d'un énoncé compère	10 mots
Taille lexique locuteur	1115 mots
Taille lexique compère	849 mots
Mots les plus utilisés par le locuteur (ordre décroissant)	Oui, je, euh, de, à, un, une, hôtel
Mots les plus utilisés par le compère (ordre décroissant)	De, à, hôtel, le, réserver, l', je

Table 1 : Mesures moyennes sur le corpus

La variabilité des énoncés produits dépend de la complexité du scénario mais également du comportement du compère. Les phénomènes les plus intéressants (références, négociations, négations…) sont observés surtout pour des scénarii complexes avec un compère non coopératif.

La Figure 3 présente un extrait de dialogue. On peut y observer les phénomènes fréquents du dialogue : hésitations, répétitions, ainsi que des références « cette nuit là » « la même chose » qui sont très nombreuses.

U : euh donc bah je voudrais réserver pour *cette nuit là* donc deux chambres à l'hôtel Mercure euh de Lille

S: je vous réserve deux chambres à l'hôtel Mercure le Grand Hôtel à Lille souhaitez-vous d'autres réservations

U: alors ça serait la *même chose* donc deux chambres également à Paris euh pour la nuit du vingt et un au vingt-deux puis du vingt-deux au vingt-trois février avec *les mêmes euh les mêmes critères* euh donc toujours deux couples avec un enfant

Figure 3 : Exemple d'échanges utilisateur-compère

4. DIFFUSION DU CORPUS

Le corpus, y compris les transcriptions et annotations sémantiques, sera diffusé par ELRA/ELDA le plus largement possible sous la forme d'une distribution qui comprendra également les résultats anonymes de l'évaluation et les outils développés pour la campagne. Le consortium prêtera attention à la réutilisabilité de ce type de ressources, afin de contribuer à la standardisation des méthodes de test. Le but de cette distribution est de permettre à un acteur externe de la campagne de s'évaluer et de comparer ses résultats à ceux produits lors de la campagne.

5. CONCLUSION

La fin de la collecte et des transcriptions du corpus MEDIA est prévue pour le mois de mars 2004. Chaque dialogue (signal et transcriptions) sera accompagné des consignes du scénario donné à l'usager d'une part, et des consignes fixant le comportement du compère d'autre part. A la suite des enregistrements, l'annotation sémantique des dialogues devrait commencer début avril. Le travail porte actuellement sur l'analyse des dialogues déjà enregistrés afin de finaliser la structure de la représentation et l'ensemble des concepts liés à la tâche.

REMERCIEMENTS

Le projet MEDIA est soutenu par l'action interministérielle française Technolangue dans le cadre de l'infrastructure d'évaluation des systèmes d'ingénierie linguistique du français EVALDA.

BIBLIOGRAPHIE

[1] J. Mariani. The Aupelf-Uref Evaluation-Based Language Engineering Action and Related Projects. In *Proceedings of the First International Conference on Language Resources and Evaluation*, volume 1, Granada, 1998

[2] J. Antoine and *al.* Predictive and objective evaluation of speech understanding: the challenge evaluation campaign of the I3 speech workgroup of the French CNRS. In *Proceedings of the third International Conference on Language Resources and Evaluation*, volume 1, Las Palmas, 2002.

[3] D. Gibbon, R. Moore and R. Winsky, *Handbook of Standards and Resources for Spoken Language Resources,* Mouton de Gruyter, New-York, 1997.

[4] L. Dybkjaer and *al.*, The Disc Approach to Spoken Language System Development and Evaluation. In *Proceedings of the First International Conference on Language Resources and Evaluation*, volume 1, Granada, 1998.

[5] E. Giachin, and S. McGlashan. Spoken Language Dialogue Systems. In S. Young and G.Bloothooft (Eds.) Corpus-based methods in language and speech processing. Dordrecht Kluwer Academic Publishers, 69-117, 1997.

[6] MADCOW. Multi-Site Data Collection for a Spoken Language Corpus, *DARPA Speech and Natural Language Workshop*, 1992.

[7] M. Walker, R. Passonneau and J. Boland. Quantitative and Qualitative Evaluation of Darpa Communicator Spoken Dialog Systems, *ACL/EACL Toulouse*, 2001.

[8] L. Devillers, H. Maynard and P. Paroubek. Méthodologies d'évaluation des systèmes de dialogue parlé: réflexions et expériences autour de la compréhension. TALN 2002.

[9] H. Maynard and L. Devillers. A framework for evaluating contextual understanding. In Proceedings of the *International Conference of Speech and Language Processing,* 2000.

[10] J. Antoine and *al.* Obtaining predictive results with an objective evaluation of spoken dialogue systems: experiments with the DCR assessment paradigm. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, volume 1, Athens, 2002.