LUNA : Compréhension en contexte pour le dialogue oral

Géraldine Damnati [1], *Frédéric Béchet* [2], *Renato de Mori* [2]

[1] France Telecom R&D - Orange Labs - 2 av. Pierre Marzin 22307 Lannion, France
[2] Université d'Avignon - LIA - 339 ch. des Meinajaries 84911 Avignon, France

ABSTRACT

This paper describes the first results achieved within the LUNA project in coupling the Spoken Language Understanding process with the Automatic Speech Recognition and Dialog Manager processes. This strategy is implemented and evaluated on a France Telecom telephone service application called FT3000.
Keywords compréhension automatique de la parole, dialogue homme-machine.

1. Introduction

La Compréhension Automatique de la Parole désigne la tâche consistant à construire une représentation du sens d'un énoncé oral. Elle peut être définie comme l'ensemble des analyses visant à caractériser, étiqueter, structurer et finalement représenter formellement l'information contenue dans un message en fonction des contextes d'élocution et d'utilisation de ceux-ci.

Le projet Européen LUNA[1] a pour objectif de développer des modèles de compréhension liés au contexte d'élocution dans le cadre du dialogue oral homme-machine. Il vise à améliorer l'efficacité des systèmes de dialogue oral homme-machine et permettre le déploiement d'applications plus complexes que celles mises en service à ce jour. Les deux axes de recherche développés sont d'une part le couplage entre les processus de Compréhension Automatique de la Parole et les processus adjacents (Reconnaissance Automatique de la Parole en amont et Gestionnaire de Dialogue en aval) ; et d'autre part le développement de stratégies de compréhension permettant de s'adapter en fonction du message et du contexte local ou global du dialogue.

Dans le cadre du projet LUNA, cet article présente les travaux menés à France Télécom et à l'Université d'Avignon sur la compréhension de la parole, sur un corpus de dialogue collecté à partir d'une application mise en service et acceptant la parole spontanée telles que les applications **10 13** et **3000** de France Télécom. Ces corpus ont deux avantages majeurs : d'une part leur quantité est quasiment illimitée, tant que l'application est en service il suffit de collecter les traces des dialogues ; d'autre part ils contiennent de *vrais* dialogues, non simulés, avec des utilisateurs novices ou expérimentés, et la parole est très spontanée.

[1] http ://www.ist-luna.eu/

2. Stratégie de compréhension dans le projet LUNA

La tâche de compréhension automatique de la parole est une tâche complexe traditionnellement décomposée en deux sous-tâches : le signal de parole est tout d'abord projeté vers un ensemble d'hypothèses lexicales ; ces hypothèses sont ensuite projetées vers les représentations conceptuelles attendues par l'application visée. En suivant un modèle largement employé, cette étape de projection des mots vers les interprétations est elle-même décomposée en deux étapes : projection dans un premier temps des mots vers des *unités sémantiques élémentaires*, souvent appelés *concepts* ; puis composition de ces concepts pour produire les interprétations complètes des énoncés. Ces concepts regroupent à la fois les actions, souvent supportées par des verbes, les objets, supportés par des groupes nominaux, et d'éventuels modifieurs. Les interprétations peuvent être représentées par des formules logiques de type *prédicat/argument* ou encore, comme préconisé dans LUNA, par des *cadres sémantiques* inspirés du modèle *FrameNet* de Berkeley. Pour chacune de ces sous-tâches le contexte du dialogue, local et global, peut être utilisé pour contraindre la recherche de la meilleure séquence de mots, de concepts et d'interprétations.

La figure 1 présente la stratégie proposée par LUNA. Dans un premier temps un module de Reconnaissance Automatique de la Parole (RAP) produit un graphe valué d'hypothèses lexicales à partir d'un message oral. Les poids des différents chemins dans ce graphe correspondent à la combinaison des scores des modèles acoustiques et de langage du système de RAP. Ce graphe de mots est projeté vers un graphe de concepts par une étape de traduction mots/concepts ; ce graphe de concepts étant à son tour projeté vers un graphe d'interprétations par un processus de composition sémantique.

Les deux principales caractéristiques de cette approche sont d'une part la prise en compte d'un ensemble d'hypothèses multiples, représentées sous la forme d'un graphe, quelle que soit le niveau de représentation (mot, concept ou interprétation) ; et d'autre part la possibilité pour chaque niveau de contraindre le graphe d'hypothèses en supprimant des chemins ou en réévaluant les chemins existant à l'aide de modèles prenant en compte le contexte *global* et *local* du dialogue. Ces modèles contextuels, comme nous

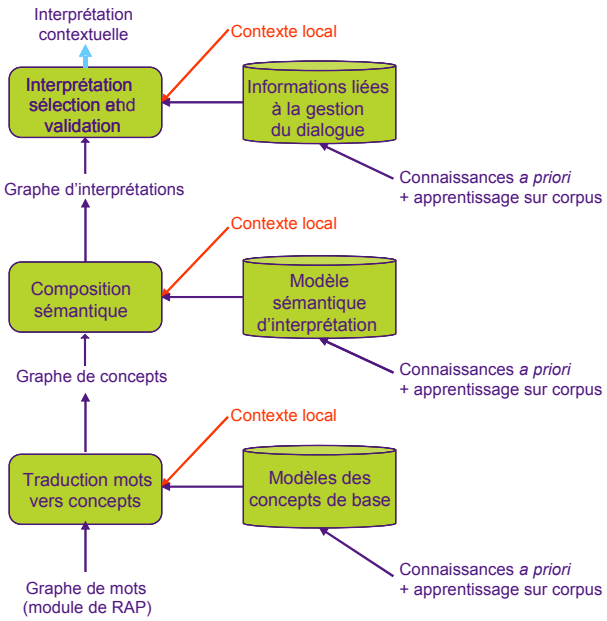

Fig. 1: Stratégie de compréhension de la parole proposée par le projet LUNA

le verrons dans le paragraphe 5, peuvent être obtenus à partir de connaissances *a priori* sur le dialogue, exprimées de manière explicite, ou bien représentés des connaissances *apprises* sur un corpus de dialogue d'entraînement.

3. Le corpus FT3000

Le service **3000** est le premier service déployé à France Télécom acceptant la parole spontanée non contrainte. Il a été mis en service en Octobre 2005. Ce service permet aux clients de France Télécom d'obtenir des renseignements, de souscrire à environ 30 services liés à leur ligne téléphonique, ou bien d'accéder à des services dédiés comme la consultation de la consommation, le paiement de la facture ou l'activation d'un transfert d'appel. Le système est fondé sur un module de reconnaissance de la parole continue avec un modèle de langage de type bigramme, le module d'interprétation sémantique de France Télécom fonctionne en deux étapes : une première étape traduit le message transcrit automatiquement en une série de concepts liés à l'application; puis un ensemble de règles logiques manuellement définies (plus de 2600 règles pour le **3000**) compose ces concepts pour produire une interprétation sous la forme prédicat/argument. Le gestionnaire de dialogue (DM) est un automate à états finis. Le DM peut emprunter plusieurs états entre deux tours de parole et les états qui engendrent un message du système sont appelés *phases* et sont au nombre de 137. Pour chaque phase est défini l'ensemble des interprétations qui permettent d'emprunter une transition dans l'automate des états de dialogue. Ainsi, la notion d'interprétation *attendue* pour une phase est donnée par la connaissance de cet ensemble. La reconnaissance d'une interprétation non-attendue entraîne une réaction d'incompréhension de la part du système.

4. Caractérisation des énoncés utilisateurs

Le corpus FT3000 contient des traces de dialogue entre un système mis en service et de vrais utilisateurs. Tous les problèmes concrets tels que les communications interrompues, le bruit dans les messages, les utilisateurs non coopératifs voire furieux, auxquels une application déployée est confrontée, se retrouvent dans ce corpus. Un corpus de 4554 énoncés collectés en mai 2007 a été validé manuellement au niveau conceptuel ainsi qu'au niveau des interprétations. Les analyses présentées dans cet article sont menées sur la base de cette référence validée manuellement. Par ailleurs un corpus de 44k énoncés collectés antérieurement a été annoté automatiquement au niveau sémantique et est utilisé comme corpus d'apprentissage pour les différents composants du système de compréhension LUNA.

Afin de mieux analyser un système de compréhension, il est important de ne pas se contenter d'évaluer les performances du système dans son ensemble mais plutôt d'observer son comportement sur les différents types de messages auxquels il est confronté. La première grande distinction concerne les messages qui contiennent une interprétation valide et ceux qui n'en contiennent pas. Dans ce dernier cas, le message doit être rejeté. Dans le corpus FT3000, 22% des messages doivent être rejetés ce qui montre que l'amélioration de la précision des systèmes de compréhension s'accompagne de la nécessité de construire des stratégies de rejet adaptées. L'ensemble des messages à rejeter est ici subdivisé en quatre catégories et l'ensemble des messages contenant une interprétation valide est quant à lui subdivisé en deux catégories. Certaines catégories relèvent d'informations contextuelles et sont définies en fonction de la phase courante associée à un message donné. Ainsi dans cet article les messages du corpus FT3000 sont classés en 6 catégories :

1. $C1$ contient les messages vides, c'est à dire contenant du bruit sans parole;
2. $C2$ contient les énoncés hors-domaine, c'est à dire ne contenant aucune information relative au service, mais plutôt des commentaires généraux ou encore des appréciations;
3. $C3$ contient les énoncés dont le thème correspond bien à celui du service, mais dont la requête n'est pas couverte par les règles d'interprétation;
4. $C4$ contient les énoncés dont la requête est couverte par les règles d'interprétation mais pour lesquels l'interprétation obtenue ne correspond à aucune interprétation *attendue* dans la phase courante.;
5. $C5$ contient les énoncés pertinents, dont les requêtes sont bien couvertes par le service, qui peuvent être pris en compte étant donnée la phase courante et considérés peu fréquents dans le contexte de cette phase courante.
6. $C6$ contient les énoncés pertinents, dont les requêtes sont bien couvertes par le service, qui peuvent être pris en compte étant donnée la phase courante et considérés fréquents dans le contexte de cette phase courante.

Le découpage entre $C5$ et $C6$ repose sur la notion de paire Phase/Interprétation fréquente et résulte de l'analyse du corpus d'apprentissage. Celuici comporte 27667 occurrences de telles paires (les énoncés à rejeter ne sont pas pris en considération ici). Sont considérées fréquentes les paires dont le nombre d'occurrences est supérieur ou égal à 100. Ce critère permet d'extraire 37 paires parmi les 1854 paires différentes observées dans le corpus d'apprentissage, et ces 37 paires couvrent 59% des occurrences. La forte dispersion de la distribution des paires Phase/Interprétation est assez prévisible dans la mesure où l'application **3000** présente quelques fonctionnalités majoritairement demandées par les utilisateurs (comme l'activation d'un transfert d'appel, le paiement de la facture...).

La table 1 indique les proportions des différentes catégories avec des exemples de messages, sur un corpus collecté sur une période de 10 jours par le service FT3000. Ce corpus représente une *photo réaliste* des messages que doit traiter une application mise en service. Comme on peut le voir, 22.3% des messages sont dans les classes $C1$, $C2$, $C3$ et $C4$ c'est à dire que 22.3% des messages devraient être rejetés car non pertinents pour le service. Ce point illustre parfaitement la différence entre les corpus extraits d'applications mises en service, et les corpus de laboratoire de type ATIS où seulement les messages des catégories $C5$ et $C6$ sont considérés. Ainsi il est primordial lors de tout transfert d'une application de laboratoire à une application industrielle d'évaluer les modèles proposés également sur des messages *a priori* à rejeter tels que les messages des catégories 1 à 4, dans la mesure où ils représentent une part non négligeable du trafic d'un système déployé.

5. Modèles contextuels

Le modèle sémantique implémentée pour le corpus FT3000 est composé de deux niveaux :

1. le premier niveau traduit une séquence de mots $W = w_1, \ldots, w_n$ en une séquence de concepts élémentaires $C = c_1, \ldots, c_m$ à l'aide de grammaires régulières représentées par un transducteur mot/concept (appelé M_1) tel que présenté dans [3] ;

2. le deuxième niveau applique les règles d'inférence sur chaque séquence de concept C afin de produire toutes les interprétations possibles pour C sous forme de structure prédicat/argument. Ces règles, au nombre d'environ 2600, sont également représentées sous forme de transducteurs, prenant en entrée les concepts et produisant les interprétations (M_2).

Ces deux transducteurs, M_1 et M_2, permettent d'imposer des contraintes sur la recherche du meilleur chemin dans le graphe d'hypothèses produit par le module de RAP. Ils représentent la connaissance *a priori* disponible sur le contexte *global* du dialogue : quels sont les objets sémantiques pertinents pour l'application visée et comment peut-on les combiner pour produire une interprétation formelle utilisable par le DM.

Contexte	*global*	*local*
a priori	M_1,M_2,M_6	M_3
appris	M_4	M_5

Tab. 2: Caractérisation des différents modèles contextuels développés pour le corpus **FT3000**

En complément de ces contraintes globales, un autre ensemble de connaissance *a priori* est utilisé : il s'agit des contraintes *locales* pour chaque phase du dialogue. En effet, le DM implémentant un automate de dialogue, seules certaines interprétations sont valides pour une phase donnée. Par exemple le concept "*oui*" dans une phase de dialogue demandant une valeur numérique. Ces contraintes locales sont également exprimée sous forme d'automates à état finis, un pour chacune des 137 phases de dialogue dans le DM du **3000**. Ce modèle est appelé $M3$ dans cette étude.

Les modèles de contraintes contextuelles présentés jusqu'ici ne réévaluent pas les hypothèses du graphe, ils se contentent de le structurer et éventuellement supprimer des chemins. Pour réévaluer les hypothèses, trois modèles numériques sont utilisés :
- M_4 : modèle statistique d'étiquetage des mots en concepts, permettant d'estimer la probabilité jointe $P(W, C)$, appris sur le corpus d'entraînement du **3000**, et basée sur une approche dérivée de [2] et présentée dans [4] ;
- M_5 : modèle basée sur les distributions de toutes les interprétations possibles pour une phase donnée, apprise sur le corpus d'entraînement ;
- M_6 : une modélisation de la priorité des règles d'inférence utilisées dans la phase de composition sémantique. Les 2600 règles d'inférence sont ordonnées par priorité, ce modèle numérique implémente cette priorité en donnant un score pour l'application de chaque règle dans le transducteur concept/interprétation.

Tous ces différents modèles sont représentés par des Automates à Etats Finis, en utilisant les outils d'AT&T *FSM Library* pour les accepteurs et les transducteurs et *GRM Library* pour le modèle de langage de l'étiqueteur de mots en concepts. Cette utilisation des automates permet une implémentation direct du modèle d'interprétation LUNA : le graphe d'hypothèses issus de la RAP est lui aussi représenté sous la forme d'un automate, l'application des différents modèles consiste uniquement à effectuer des opérations de composition ou d'intersection entre les automates. Plus de détails sur cette implémentation peuvent être trouvé dans [1].

La tableau 2 récapitule ces différents modèles. En fonction de l'utilisation, dans le processus d'interprétation, d'un ou plusieurs de ceux-ci, on peut contraindre la recherche de la meilleure hypothèse d'interprétation avec différents contextes : les modèles définis *a priori* définissent l'*acceptabilité* d'une hypothèse, globalement ou dans la phase courante de dialogue ; les modèles appris caractérisent sa *plausibilité* par rapport à l'utilisation courante de l'application. En comparant les résultats obtenus avec différentes configurations, on peut caractériser chaque message vocal selon les différentes catégories présentées dans le paragraphe 4, et appliquer des stratégies de valida-

Cat.	nb	%	exemple
$C1$	246	5.4%	*biiiiiip*
$C2$	231	5.1%	*bon bon qu'est ce que je dois dire là euh euh*
$C3$	472	10.4%	*on m'appelle tout le temps y'a personne je sais pas qui c'est*
$C4$	64	1.4%	*oui* en réponse à une question ouverte
$C5$	1870	41.0%	*j' appelle pour avoir un renseignement euh pour euh avoir obtenir le service du réveil*
$C6$	1671	36.7%	*je voudrais payer ma facture*

Tab. 1: Différentes catégories de messages sur le corpus FT3000

tion différentes.

6. Expériences

L'évaluation du module de compréhension de la parole est réalisée en termes de taux d'interprétation correct. Les interprétations produites dans le cadre de l'application FT3000 sont représentées par une composition de paires attribut-valeurs (ex : *Gest(Desactiver,TransfertAppel)*). Ainsi pour les catégories $C5$ et $C6$, une interprétation est considérée correcte si tous les éléments qui la composent sont corrects. Pour les 4 premières catégories en revanche un énoncé correct est un énoncé rejeté.

% correct modèle context.	Rejet C1-C4	Int. peu freq C5	Int. freq. C6
aucun	**64.9**	83.6	93.0
$M_{1,2} + M_6$	52.4	84.3	93.8
$M_{1,2,3} + M_6$	52.4	84.4	93.8
$M_{1,2,3,4} + M_6$	52.4	**85.0**	94.8
$M_{1,2,3,4,5,6}$	52.4	82.0	**94.9**

Tab. 3: Comportement des modèles contextuels en terme d'interprétation correcte pour chaque catégorie de messages

Afin d'illustrer le comportement de chacun des modèles contextuels en fonction des différentes catégories d'énoncés, le taux d'énoncés correctement reconnus relativement au nombre total de chaque catégorie est donné dans le tableau 3. La ligne *aucun* signifie qu'aucun modèle contextuel n'a été utilisé pour produire la séquence de mots sur laquelle le module de compréhension du système **3000** va extraire une interprétation : seuls les modèles de RAP sont utilisés. Les autres lignes indiquent quels modèles contextuels ont été utilisés, selon la stratégie LUNA, pour produire cette interprétation.

La première remarque concerne les écarts importants en termes de performance des modèles pour les différentes catégories de messages. Les messages les moins bien traités étant les messages hors-domaine, les messages du domaine non-couverts et les messages non-attendus. Les interprétations fréquentes présentent un taux de bonne détection important.

De façon générale, l'utilisation du contexte permet d'améliorer le taux d'interprétations correctes pour les messages contenant une interprétation valide. L'apport du contexte *appris* (M_4 et M_5) n'est perceptible en revanche que pour les paires Phase/Interprétations fréquentes ce qui se comprend aisément compte tenu des limitations des techniques d'apprentissage face aux événements rares. Mais ceci se fait au détriment du taux de détection des rejets comme le montrent les performances obtenues sur les 4 premières catégories. En effet l'utilisation des modèles d'interprétation M_1 et M_2 sur le graphe d'hypothèses lexicale issu de la RAP a tendance à chercher systématiquement un chemin produisant une interprétation valide, ce qui a pour effet de diminuer le taux de faux rejets dans C3 et C4, mais augmente le taux de fausse alarme dans C1, C2, C3 et C4. Les différences de comportement de chacun des modèles en fonction des différentes catégories de messages incitent à les envisager non pas séparément comme plusieurs méthodes alternatives mais dans leur complémentarité.

7. Conclusion

Nous avons montré dans cette étude un exemple d'implémentation de la stratégie de compréhension de la parole proposée par le projet LUNA sur une application déployée par France Télécom. La grande variabilité des messages devant être traité fait qu'il est intéressant de développer des stratégies adaptées à chaque type de message. Nous avons proposé dans cet article différents types de modèles intégrant des connaissances contextuelles (locales et globales, connues *a priori* ou apprises) dans le processus de recherche de la meilleure interprétation. Nous avons montré que ces modèles offraient des performances différentes selon le type de messages traités, et qu'ils peuvent être à la base de stratégies d'interprétation employant différentes combinaisons de modèles. C'est sur la définition de telles stratégies que nos efforts se portent actuellement.

Références

[1] Géraldine Damnati, Frédéric Béchet, and Renato de Mori. Spoken language understanding strategies on the France Telecom 3000 Voice Agency corpus. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Honolulu, HI, April 2007.

[2] E. Levin and R. Pieraccini. Concept-based spontaneous speech understanding system. Madrid, Spain, 1995.

[3] Christian Raymond, Frederic Bechet, Renato De Mori, and Geraldine Damnati. On the use of finite state transducers for semantic interpretation. *Speech Communication*, 48,3-4 :288–304, 2006.

[4] Christophe Servan, Christian Raymond, Frederic Bechet, and Pascal Nocera. Conceptual decoding from word lattices : Application to the spoken dialogue corpus MEDIA. pages 1614–1617, Pittsburgh, PA, USA, 2006.