Décodage conceptuel à partir de graphes de mots sur le corpus de dialogue Homme-Machine MEDIA

Christophe Servan, Christian Raymond, Frédéric Béchet, Pascal Nocéra

LIA - Université d'Avignon, BP1228 84911 Avignon cedex 09 France
{christophe.servan,christian.raymond,frederic.bechet,pascal.nocera}@univ-avignon.fr

ABSTRACT

Within the framework of the French evaluation program MEDIA on spoken dialogue systems, this paper presents the methods proposed at the LIA for the robust extraction of basic conceptual constituents (or concepts) from an audio message. The conceptual decoding model proposed follows a stochastic paradigm and is directly integrated into the Automatic Speech Recognition (ASR) process. This approach allows us to keep the probabilistic search space on sequences of words produced by the ASR module and to project it to a probabilistic search space of sequences of concepts. The experiments carried on on the MEDIA corpus show that the performance reached by our approach is better than the traditional sequential approach that looks first for the best sequence of words before looking for the best sequence of concepts.

1. INTRODUCTION

Dans le cadre des applications de dialogue homme-machine, la campagne MEDIA [1] s'est focalisée sur l'évaluation de systèmes de décodage conceptuel permettant d'associer à une séquence de mots une séquence de concepts relatifs au type de dialogue visé. Cette évaluation a été faite sur des transcriptions manuelles de messages audio obtenus grâce à un protocole de type *Magicien d'Oz* sur une tâche de réservation hôtelière. En complément de la campagne MEDIA, cette étude présente les premiers travaux effectués sur le corpus audio MEDIA. En utilisant d'une part le système de Reconnaissance Automatique de la Parole (RAP) SPEERAL [3] et d'autre part le module d'interprétation sémantique développé au LIA [4] nous montrons comment une approche de décodage *intégrée* cherchant directement la meilleure séquence de concepts à partir d'un graphe de mots issus du module de RAP surpasse l'approche séquentielle traditionnelle consistant à détecter les concepts uniquement dans la meilleure hypothèse de phrase sortie par le module de RAP.

Cet article est structuré comme suit : le paragraphe 2 introduit rapidement le corpus MEDIA ; les modèles de RAP développés sur ce corpus sont présentés dans le paragraphe 3 ; le paragraphe 4 rappelle brièvement l'approche du LIA concernant l'interprétation sémantique de message audio ; enfin la partie 5 décrit les résultats des expérimentations effectuées en essayant de répondre aux deux questions suivantes :

– quel est l'impact du taux d'erreurs mots sur les performances du module de décodage conceptuel ?
– quelles sont les différences de performance constatées entre d'une part l'approche de décodage *intégré* proposée dans cette étude et d'autre part l'approche séquentielle traditionnelle.

2. LE CORPUS MEDIA

La campagne d'évaluation MEDIA [1] (programme Technolangue/Evalda) se place dans le cadre de la simulation d'un système d'accès à des informations touristiques et des réservations d'hôtel. Un corpus de 1250 dialogues a été enregistré par ELDA selon un protocole de type *Magicien d'Oz* : 250 locuteurs ont effectué chacun 5 scénarios de réservation d'hôtel avec un système de dialogue simulé par un opérateur humain. Ce corpus a ensuite été transcrit manuellement, puis annoté sémantiquement selon un dictionnaire sémantique de concepts mis au point par les partenaires du projet. Le dictionnaire sémantique utilisé pour annoter le corpus MEDIA permet d'associer 3 types d'information à un mot ou un groupe de mots :

– tout d'abord une paire attribut-valeur, correspondant à une représentation sémantique *à plat* d'un énoncé ;
– puis un spécifieur qui permet de définir des relations entre les attributs et qui par conséquent peut être utilisé pour construire une représentation hiérarchique de l'interprétation d'un énoncé ;
– enfin une information sur le *mode* attaché à un concept (positif, affirmatif, interrogatif ou optionnel).

n	W^{c_n}	c_n	*valeur*
0	euh	null	
1	oui	réponse	oui
2	l'	LienRef-coRef	singulier
3	hôtel	BDObjet	hotel
4	dont	null	
5	le prix	objet	paiement-montant
6	ne dépasse pas	comparatif-paiement	inferieur
7	cent dix	paiement-montant-ent	110
8	euros	paiement-devise	euro

TAB. 1: Exemple de message annoté du corpus MEDIA

La table 1 présente un exemple de message annoté du corpus. La première colonne correspond au numéro du segment dans le message, la deuxième colonne à la chaîne de mots W^{c_n} porteuse du concept c_n contenu dans la troisième colonne. La quatrième colonne contient la valeur du concept c_n dans la chaîne W^{c_n}. Le dictionnaire sémantique MEDIA contient 83 attributs, auxquels peuvent s'ajouter 19 spécifieurs de relations entre attributs. Le corpus collecté a été découpé en plusieurs lots. Nous utilisons dans cette étude les 4 premiers lots comme corpus d'apprentissage, soit 720 dialogues contenant environ 12K messages, et le lot *Test à blanc* comme corpus de tests contenant 79 dialogues avec 1.3K messages.

3. DÉVELOPPEMENT D'UN SYSTÈME DE RAP SUR LE CORPUS MEDIA

3.1. Apprentissage des modèles

Le décodeur SPEERAL [3] a été utilisé pour transcrire les messages du corpus MEDIA. Ces messages sont enregistrés dans des conditions identiques à celles que l'on peut trouver dans un système mis en service. Les utilisateurs ont effectué leurs appels depuis leur téléphone, fixe ou cellulaire, et la qualité des enregistrements est variable. Les modèles acoustiques téléphoniques utilisés sont ceux développés lors de la campagne d'évaluation ESTER sur la transcription de données radiophoniques, ils ont ensuite été adaptés sur les 720 dialogues des lots 1,2,3,4 du corpus MEDIA par une adaptation de type MAP. Ce sont des modèles triphones.

Le modèle de langage, de type 3-grammes de mots, a été appris sur un corpus extrait des transcriptions manuelles des lots 1,2,3,4. Ce corpus contient un ensemble de 226K mots. Un lexique de 2028 mots a été défini sur ce corpus, il a été phonétisé avec l'outil LIA_PHON[1]. Sur le corpus de test utilisé (lot *Test à blanc* du corpus MEDIA), le taux de mots hors-vocabulaire du lexique choisi est de 1,6%. La perplexité est de 26,5.

Le taux d'erreur mot (ou *Word Error Rate WER*) de la transcription automatique du lot *Test à blanc* avec les modèles présentés (combinaison des modèles acoustiques et linguistiques dans le décodeur SPEERAL) est de 32,2%.

3.2. Graphes de mots

L'approche intégrée de décodage conceptuelle défendue dans cette étude nécessite le traitement de graphes de mots issus du module de RAP. Ces graphes sont produits par le décodeur SPEERAL, toutes les opérations sur les graphes sont ensuite effectuées avec l'ensemble d'outils de manipulation d'automates *AT&T FSM/GRM Library* [2]. Ces graphes nous permettent également de faire varier le WER de la meilleure hypothèse produite par le module de RAP. En effet, un but de cette étude est d'étudier la corrélation entre le taux d'erreur sur les mots et celui sur les concepts. Il est donc intéressant de produire des sorties multiples. Ces sorties sont obtenues par la méthode suivante :

- tout d'abord des graphes de mots sont générés par SPEERAL sur le corpus de test avec les modèles présentés précédemment ; en prenant la meilleure séquence de mots dans ces graphes nous obtenons les hypothèses *baseline* avec un WER moyen de 32,2% ;
- un nouveau modèle de langage (toujours de type 3-grammes) est alors appris, cette fois sur le corpus de test ;
- ce modèle est appliqué aux graphes de mots préalablement produits, après une interpolation avec le modèle *baseline* appris sur le corpus d'apprentissage ;
- en faisant varier le coefficient d'interpolation, on peut faire varier le taux d'erreur mots.

Avec cette méthode nous avons obtenu 4 décodages différents de notre corpus de test obtenus avec 4 valeurs différentes du coefficient d'interpolation (0,0 0,5 0,8 et 1.0). Ces décodages sont représentés par 4 séries de graphes de mots $G_{0,0}$, $G_{0,5}$, $G_{0,8}$ et $G_{1,0}$ dont les scores sont une combinaison des modèles acoustique et de langage. Les graphes $G_{0,0}$ correspondent au décodage *baseline* où aucune données de test n'est intégrée dans l'apprentissage. Les taux d'erreurs mots des meilleures hypothèses de ces graphes sont :

Graphes	$G_{0,0}$	$G_{0,5}$	$G_{0,8}$	$G_{1,0}$
WER	32,2	27,2	24,1	18,5

Même si l'introduction de données de tests dans l'apprentissage génère forcément un biais, il est réduit du fait que cette introduction n'intervient que dans la deuxième passe de l'étape de reconnaissance : les erreurs et confusions acoustiques produites par le modèle *baseline* sont toujours présentes. Cependant ce sont bien évidemment les résultats obtenus sur les graphes $G_{0,0}$ qui sont les plus réalistes puisqu'ils sont produits sans introduction des données de tests. Les autres graphes ne servent qu'à observer la corrélation taux d'erreur mots et taux d'erreur concepts.

4. STRATÉGIE D'INTERPRÉTATION

Nous noterons C l'interprétation d'un message. C représente une séquence de concepts de base, tels que ceux définis dans le corpus MEDIA et présenté dans l'exemple de la table 1. Le décodage conceptuel consiste à chercher la chaîne de concepts $C = c_1, c_2, \ldots, c_k$ maximisant $P(C|A)$, A étant la séquence d'observations acoustiques. Trouver la meilleure séquence de concepts $\hat{C}$ exprimée par la séquence de mots W à partir de la séquence d'observations acoustiques A s'exprime par la formule suivante :

$$P(\hat{C}\hat{W}|A) \approx \max_{C,W} P(A|W)P(W,C) \qquad (1)$$

Les deux stratégies possibles pour obtenir $\hat{C}$ sont :

- chercher tout d'abord la meilleure chaîne de mots $\hat{W}$ étant donné A, puis chercher la meilleure séquence de concepts $\hat{C}$ sur la chaîne $\hat{W}$; nous appellerons cette approche l'approche *séquentielle* ;
- chercher conjointement la meilleure chaîne de mots et la meilleure séquence de concepts, tel que cela est exprimé dans l'équation 1 ; c'est l'approche *intégrée* proposée dans cette étude.

Cette recherche par l'approche intégrée de la meilleure interprétation $\hat{C}$ est faite dans un graphe de mots produit par le système de RAP pour chaque message traité. La première étape dans cette recherche consiste à transformer ce graphe de mots en un graphe de concepts. Le principe général de cette approche est décrit dans [4], son utilisation dans la campagne MEDIA est présentée dans [5], nous allons brièvement la résumer dans le paragraphe suivant.

Les constituants sémantiques sont appelés *tags conceptuels* et sont notés *c*. Ils correspondent aux 83 attributs présentés dans l'ontologie MEDIA (les informations sur les spécifieurs et les modes sont associés à un autre niveau d'interprétation dans notre système). À chaque tag c est associée la chaîne de mot W^c supportant le concept et à partir de laquelle sa valeur va être extraite, comme dans l'exemple de la table 1.

Dans le module de compréhension développé au LIA il existe un automate à états finis (ou Finite State Machine *FSM*) pour chacun de ces concepts. Ces automates sont

[1] téléchargeable à l'adresse :
http ://www.lia.univ-avignon.fr/chercheurs/bechet/

des transducteurs qui acceptent les séquences de mots W^c en entrée et qui produisent en sortie les concepts c correspondant. Ces transducteurs peuvent être créés manuellement pour les concepts indépendants du domaine (par exemple les dates ou les prix), ou induits par apprentissage pour les concepts propres au corpus MEDIA. L'ensemble de ces transducteurs est regroupé en un seul automate appelé *automate conceptuel*, auquel est ajouté un automate *filler* pour accepter tout ce qui ne fait pas partie d'un concept.

En effectuant une opération d'intersection entre le graphe de mots produit par le système de RAP et cet *automate conceptuel*, nous obtenons directement un transducteur où les chemins sur les symboles d'entrées sont des chaînes de mots et les chemins sur les symboles de sorties sont des chaînes de concepts. Afin d'évaluer les différentes analyses possibles en concepts d'une même chaîne de mots, un étiqueteur en concepts à base de HMM, lui aussi représenté sous la forme d'un automate, est composé avec le transducteur obtenu.

Le résultat du processus de décodage est une liste de n-meilleures interprétations appelée *N-Best Structurée*. Cette liste contient les meilleures interprétations du transducteur final structurées selon deux niveaux : le premier niveau correspond aux meilleures chaînes de concepts ; le deuxième niveau contient pour chaque séquence de concepts les meilleures valeurs trouvées dans le transducteur. La dernière étape du processus d'interprétation réside dans le module de décision, basé sur des classifieurs, choisissant une hypothèse dans cette liste de n-meilleure hypothèses. C'est à cette étape que le contexte du dialogue peut intervenir. Dans les expériences présentées dans le paragraphe 5, le module de décision est réduit au choix de l'hypothèse de probabilité maximale dans le transducteur de décodage. Cette stratégie d'interprétation est présentée à la figure 4.

Nous appellerons cette méthode l'approche *intégrée*, dans la mesure ou la recherche de la meilleure chaîne de mots et de la meilleure chaîne de concepts est simultanée. Pour comparer cette approche à l'approche séquentielle traditionnelle, nous avons également fait les mêmes expériences en réduisant le graphe de mots produit par SPEERAL à la chaîne de mots de probabilité maximale selon les modèles de RAP. La chaîne de traitement est par la suite identique.

5. EXPÉRIENCES

Les expériences présentées dans cette étude ont été menées sur le corpus MEDIA en considérant les 83 attributs présentés au paragraphe 2. Le mode et les 19 spécifieurs ne sont pas pris en compte ici, ils sont traités dans notre système par le module d'interprétation d'un énoncé en contexte, et ne relèvent donc pas du processus de décodage conceptuel présenté ici. Les performances sont mesurées par rapport au taux d'erreurs sur les paires attribut/valeur (appelé le *Concept Error Rate* ou *CER*, cette mesure est obtenue de manière similaire au WER en considérant les concepts à la place des mots). Un concept détecté est considéré comme correct uniquement si l'attribut du concept ainsi que sa valeur normalisée sont corrects d'après la référence.

Le tableau 2 présente les résultats des deux approches, *séquentielle* et *intégrée* sur plusieurs graphes de mots. Nous observons que dans tous les cas l'approche intégrée surpasse l'approche séquentielle, sauf bien sûr pour les transcriptions manuelles où le graphe de mots est réduit à une seule chaîne. Ainsi l'approche intégrée permet d'obtenir des performances similaires à ce que l'on obtiendrait en séquentiel avec un taux d'erreurs mots inférieur de 15% en relatif (cf écart entre $G_{0,0}$ et $G_{0,5}$).

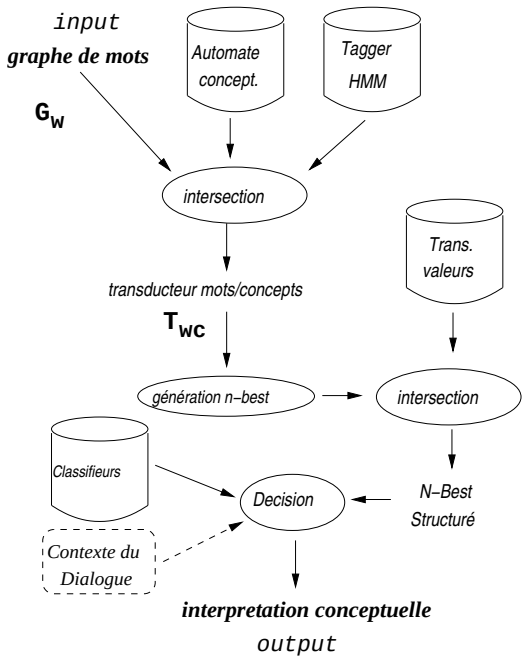

FIG. 1: Stratégie de compréhension du langage naturel oral du LIA

Il faut noter également que les performances obtenues sur les transcriptions manuelles sont comparables à celles des meilleurs systèmes ayant participés à l'évaluation MEDIA.

Un autre enseignement intéressant de ces expériences est la corrélation entre le taux d'erreur sur les mots (WER) et celui sur les concepts (CER). La figure 2 illustre cela en montrant une relation linéaire entre ces deux quantités.

Graphe	$G_{0,0}$	$G_{0,5}$	$G_{0,8}$	$G_{1,0}$	Ref.
WER	32.2	27.2	24.1	18.5	0
CER (Seq.)	44.8	41.2	39.3	36.5	20.9
CER (Int.)	40.8	38.5	37.7	34.2	20.9

TAB. 2: WER et CER sur différents graphes avec l'approche séquentielle (Seq.) et l'approche intégrée (Int.). La colonne *Ref.* correspond au traitement de la transcription manuelle du corpus de test

La dernière expérience présentée dans cette étude concerne l'évaluation des listes de n-meilleures hypothèses produites par les différentes méthodes testées. Ces listes sont particulièrement intéressantes dans le cadre d'un dialogue car il est possible de fournir au gestionnaire de dialogue, non pas une hypothèse unique, mais plusieurs hypothèses que le contexte du dialogue peut aider à filtrer. Une mesure communément utilisée pour mesurer le potentiel d'un graphe ou d'une liste d'hypothèses est la mesure *Oracle*.

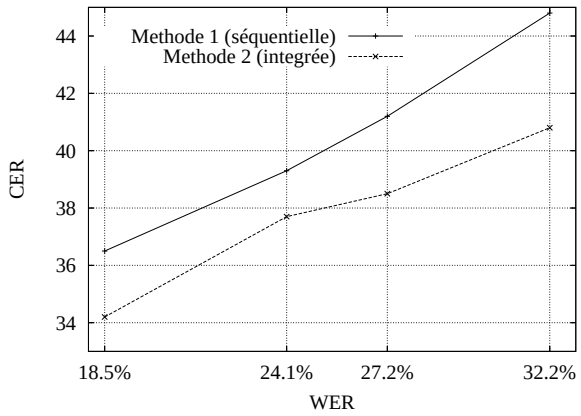

FIG. 2: Évolution du CER en fonction du WER

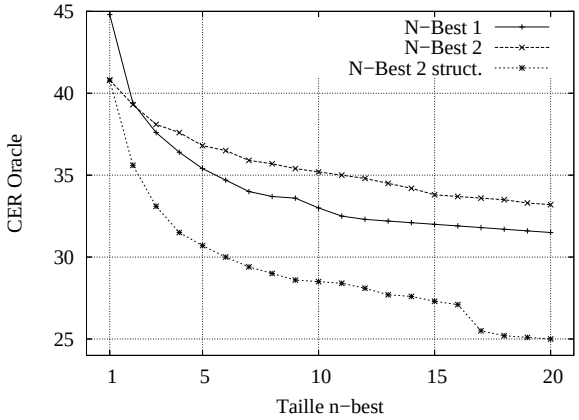

FIG. 3: Évolution du CER Oracle en fonction des tailles des listes de n-meilleures hypothèses pour deux méthodes : N-Best 1 = méthode séquentielle et N-Best 2 = méthode intégrée (avec et sans liste structurée)

Cette mesure consiste à sélectionner dans un ensemble d'hypothèses celle qui a le plus petit taux d'erreurs. Elle constitue ainsi le taux d'erreur minimal que ferait un système qui prendrait toujours la bonne décision sur le filtrage d'une liste d'hypothèses. Trois listes d'hypothèses ont été produites à partir des graphes $G_{0,0}$ (graphes *baseline* n'incluant pas de corpus de test dans l'apprentissage des modèles), leurs évaluations sont présentées sur la figure 3 :

- *N-Best 1* : cette liste est obtenue en énumérant la liste des n-meilleures hypothèses obtenues avec la méthode séquentielle ; la chaîne de mots étant fixe, chaque hypothèse diffère au niveau de la liste des concepts ;
- *N-Best 2* : cette fois les n-meilleures hypothèses sont les n-meilleures chemins dans le transducteur générées avec la méthode intégrée ; la chaîne de mot n'étant pas fixe, les n-meilleures chemins contiennent souvent la même suite de concepts et ne varie que par des choix de mots différents ;
- *N-Best 2 struct.* : correspond au N-Best Structuré décrit dans [4], et obtenu avec la méthode intégrée.

Comme le montre la figure 3, le N-Best Structuré permet d'éviter le principal inconvénient des listes de n-meilleures hypothèses produites à partir de graphes de mots : la génération d'hypothèses qui ne diffèrent que par des mots non signifiant du point de vue de l'interprétation du message.

En structurant cette liste par chaînes de concepts et valeurs, on obtient un résumé de toutes les interprétations possibles contenues dans le graphe en un nombre restreint d'hypothèses. Par exemple, en ne gardant que les 3 meilleures hypothèses du N-Best Structuré, on obtient le même taux Oracle qu'avec la liste complète des 20 meilleures hypothèses des autres méthodes.

6. CONCLUSION

Nous avons présenté dans cette étude un modèle de décodage conceptuel, basé sur une approche stochastique, intégré directement dans le processus de Reconnaissance Automatique de la Parole (RAP). L'un des principaux avantages de cette approche est de garder l'espace probabiliste des phrases produit en sortie du module de RAP et de le projeter vers un espace probabiliste de séquences de concepts. Ainsi l'incertitude dans l'interprétation d'un message peut-elle être gardée plus longtemps pour être levée par des niveaux supérieurs d'interprétation intégrant le contexte du dialogue.

Les expériences menées sur le corpus MEDIA montrent que les performances du décodage conceptuel se dégradent linéairement en fonction du taux d'erreurs sur les mots. Nous avons cependant montré qu'une approche *intégrée* cherchant conjointement la meilleure séquence de mots et de concepts donnait de meilleurs résultats qu'une approche séquentielle.

Enfin la génération d'une liste de n-meilleures hypothèses structurées permet de réduire considérablement le nombre d'hypothèses susceptibles d'être envoyées au gestionnaire de dialogue, en gardant le même taux d'erreurs Oracle que la liste complète.

RÉFÉRENCES

[1] Helene Bonneau-Maynard, Sophie Rosset, Christelle Ayache, Anne Kuhn, and Djamel Mostefa. Semantic annotation of the french media dialog corpus. In *Proceedings of Eurospeech*, Lisboa, Portugal, 2005.

[2] Mehryar Mohri, Fernando Pereira, and Michael Riley. Weighted finite-state transducers in speech recognition. *Computer, Speech and Language*, 16(1) :69–88, 2002.

[3] P. Nocera, G. Linares, and D. Massonie. Principes et performances du décodeur parole continue Speeral. In *Proc. Journées d'Etude sur la Parole (JEP)*, 2002.

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

[5] Christophe Servan and Frederic Bechet. Décodage conceptuel et apprentissage automatique : application au corpus de dialogue homme-machine media. In *TALN*, Leuven, 2006.