Utilisation des transducteurs dans le décodage conceptuel : application au corpus MEDIA

Christophe Servan

Utilisation des transducteurs dans le décodage conceptuel : application au corpus MEDIA

Christophe Servan
LIA - Université d'Avignon
christophe.servan@univ-avignon.fr

Résumé : Cet article présente les travaux du LIA effectués sur le corpus MEDIA visant à utiliser des méthodes statistiques dans la compréhension de la parole spontanée ; c'est-à-dire, l'extraction, à partir d'un message audio, d'une séquence de concepts élémentaires. Le modèle de décodage conceptuel présenté est basé sur une approche stochastique qui utilise des transducteurs permettant d'intégrer directement le processus de compréhension au processus de Reconnaissance Automatique de la Parole (RAP). Cette approche permet 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.
Mots-clés : Dialogue Homme-Machine, Reconnaissance Automatique de la Parole, Apprentissage Automatique à base de corpus

1 INTRODUCTION

Dans les applications de dialogue homme-machine téléphonique, le processus d'interprétation consiste à extraire du message oral des structures conceptuelles. Cette opération ne se résume pas forcément à une analyse de la transcription textuelle du message par une grammaire syntaxico sémantique. Plusieurs considérations étayent cette proposition : d'une part les règles d'interprétation peuvent être contextuelles ; d'autre part, dans le traitement de la parole spontanée, des parties entières du message peuvent être inutiles à la compréhension de celui-ci et l'opération de reconnaissance de concepts peut être un succès même si l'ensemble du message n'est pas complètement engendré par une grammaire. Enfin, la même séquence de mots peut générer différentes interprétations.

Plusieurs formalismes ont été proposés pour décrire des structures sémantiques. Ils sont essentiellement basés sur les concepts d'entités et de relations. En général les concepts généraux représentant l'interprétation complète d'un message sont obtenus par des opérations de composition sur des concepts élémentaires. Ces concepts sont relativement indépendants du modèle sémantique global utilisé. Ils représentent à la fois les objets sémantiques manipulés par l'application, correspondant à des catégories d'entités nommées telles que les dates, les prix, ou encore les noms propres (ville, hôtel, ...) ; mais aussi les actes dialogiques. C'est sur ces concepts élémentaires que s'est focalisée la campagne d'évaluation MEDIA (programme Technolangue/Evalda), qui consiste à évaluer les capacités d'interprétations de plusieurs systèmes sur un corpus de traces de dialogue homme-machine portant sur un serveur d'informations touristiques.

Cette étude présente les travaux du LIA effectués sur le corpus MEDIA, visant à proposer des méthodes d'analyse robuste permettant d'extraire d'un message audio une séquence de concepts élémentaires. Ces concepts sont les entités utilisées pour construire une interprétation sémantique complète des messages traités. La campagne MEDIA est structurée en deux phases : une phase d'évaluation de la compréhension *hors contexte* et une autre *en contexte*. Dans la première, les énoncés sont traités indépendamment les uns des autres, sans aucune information sur le dialogue en cours. Dans la deuxième, les concepts sont enrichis avec les informations contextuelles obtenues lors des précédents tours de dialogue. Nous nous focaliserons dans cette étude sur l'interprétation *hors contexte* des concepts élémentaires.

Ce papier est organisé comme suit : après avoir rapidement présenté la problématique du décodage conceptuel dans le cadre du projet MEDIA, nous présenterons l'approche privilégiée par de récents travaux au LIA [Servan and Bechet, 2006] pour résoudre ce problème, celle qui envisage ce processus comme un processus de traduction automatique. Le paragraphe 2.4 présente le cadre théorique de cette étude puis les différents composants de l'approche proposée sont détaillés dans les paragraphes 3. Nous présenterons le fonctionnement général du système dans le paragraphe 5 dont les détails seront présenté au paragraphes 4.1 et 4.2. Enfin le paragraphe 6 présentera les résultats obtenus par notre approche sur le corpus MEDIA.

2 DÉCODAGE CONCEPTUEL POUR LES SYSTÈMES DE DIALOGUE

2.1 Compréhension dans les systèmes de dialogue

Les applications de dialogue homme-machine considérées dans cette étude peuvent être vues comme une interface entre un utilisateur et une base de données. Le but du dialogue est de remplir tous les champs d'une requête qui va être adressée à la base de données. Dans ce cadre les concepts sémantiques de base sont de 3 types : les concepts relatifs au type de la requête ; les

n	W^{c_n}	c_n	*mode*	*spécifieur*	*valeur*
0	alors je voudrais réserver	command-tache	+		reservation
1	pour le vingt sept	temps-date	+	reservation	27/12
2	vingt huit	temps-date	+	reservation	28/12
3	et vingt neuf décembre	temps-date	+	reservation	29/12
4	entre Narbonne	localisation-ville	+	hotel-debut	narbonne
5	et Limoges	localisation-ville	+	hotel-fin	limoges
6	pour un couple	sejour-nbCouple	+	reservation	1
7	et un enfant	sejour-nbEnfant	+	reservation	1
8	comptez au niveau	null	+		
9	des prix	objet	+	reservation-chambre	paiement-montant-entier
10	soixante	paiement-montant-entier	+	reservation-chambre	60
11	euros	paiement-monnaie	+		euro
12	par chambre	null	+		

TAB. 1 – Exemple de message annoté du corpus MEDIA

concepts relatifs aux valeurs qui instancient les paramètres de la requête ; et enfin les concepts relatifs à la conduite du dialogue. La campagne d'évaluation MEDIA [MEDIA, 2005] (programme Technolangue/Evalda) se place dans ce cadre applicatif à travers 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 *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 MEDIA [MEDIA, 2005]. Ce corpus est décrit brièvement dans le prochain paragraphe.

2.2 Le corpus MEDIA

Le dictionnaire sémantique utilisé pour annoter le corpus MEDIA [MEDIA, 2005] 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).

La table 1 présente un exemple de message annoté du corpus MEDIA. 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. Les colonnes 4, 5 et 6 contiennent le mode, le spécifieur et 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 d'utilisateurs, et le lot 5 comme corpus de tests contenant 200 dialogues avec 3K messages d'utilisateurs.

2.3 Décodage conceptuel stochastique

Le traitement automatique de messages oraux a deux particularités : d'une part l'ensemble des caractéristiques de la parole spontanée inhérente aux "disfluences" (répétitions, reprises, inscises, hésitations, *etc.*) ; d'autre part une destructuration de la phrase prononcée, caractérisée par une absence de toute ponctuation et de segmentation, à l'exception des silences du locuteur. Il est à noter que ces silences sont notés en fonction d'un seuil fixé. Ces deux particularités rendent les analyses complètes de ce type de message très difficiles, ce qui oblige la plupart des systèmes à se tourner vers une analyse partielle du message. Nous savons qu'en dehors du fait qu'il existe deux grandes familles d'analyse (à base de connaissance et à apprentissage), le décodage conceptuel peut être vu suivant deux méthodes différentes :

– comme une extension d'un décodage syntaxique, auquel va venir s'ajouter l'information sémantique. Dans ce cas là, décoder le message peut s'assimiler à une structure ou un ensemble de structures dont les nœuds sont les concepts ;
– comme un processus de traduction automatique où le flux d'entrée (les mot) va donner un graphe de symboles en sortie (les concepts).

Le système TINA [Seneff, 1992] est une bonne illustration de la première méthode, tout comme le système du LORIA [Denis and Al., 2006], utilisé lors de la campagne MEDIA, qui illustre un décodage conceptuel complet séquentiel à partir d'une analyse morpho-syntaxique. Ce dernier est un système à base de règles de grammaire formelle.

La seconde méthode se rapproche du domaine de la Reconnaissance Automatique de la Parole (RAP). Dans ce domaine, le traitement de la parole est vu comme la transmission d'un signal dans un canal bruité. Le but est d'analyser le message à partir des observations (les paramètres acoustiques dans la RAP, les mots dans le décodage conceptuel) qui sont passés à travers le canal de communication. Cette opération est assimilable à un système de traduction, qui se réalise de manière statistique stochastique en maximisant $P(C|A)$, A étant la séquence acoustique, et C l'interprétation à trouver. Ces travaux ont été initiés par [Levin and Al., 1995], et se retrouve dans de

nombreux systèmes de décodage conceptuel tels que celui du LIMSI [Maynard and Al., 2005].

2.4 Modèle théorique

Le modèle théorique developpé dans cette étude appartient à la deuxième famille d'analyse présentée dans le paragraphe précédent. Ce choix a été fait pour deux principales raisons :

– d'une part, les modèles utilisés par un système de RAP prennant en compte un historique faible, il apparait interessant de considérer l'ensemble de solutions possibles sous forme de graphe de mots ;
– d'autre part, malgré le fait que les analyse syntaxiques soient applicables à un graphe de mots, il reste cependant les problèmes inhérents au langage naturel (disfluences, absence de structure, *etc.*) ; ces difficultés rendent ce type d'analyse particulièrement difficiles et rare sont celles qui arrivent à produire une analyse complète du message.

Nous noterons C l'interprétation d'un message. C représente une séquence de concepts de base, tels que ceux définis dans le corpus MEDIA. L'approche la plus utilisée est un décodage séquentiel cherchant tout d'abord à maximiser la meilleure chaîne de mots $\hat{W}$, puis de rechercher la meilleure interpréation $\hat{C}$ sachant $\hat{W}$:

$$\begin{aligned} P(\hat{W}) &\approx \underset{W}{\textit{argmax}}\ P(A|W)P(W) \\ P(\hat{C}) &\approx \underset{C}{\textit{argmax}}\ P(C|\hat{W}) \end{aligned} \quad (1)$$

L'inconvénient principal de cette méthode est le choix de $\hat{W}$ uniquement à partir d'un n-gramme de mots.

Notre approche du 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. En utilisant le même paradigme que celui utilisé en RAP, trouver la meilleure séquence de concepts $\hat{C}$ exprimé par la séquence de mots W à partir de la séquence d'observation acoustique A s'exprime par la formule suivante :

$$P(\hat{C}, \hat{W}|A) \approx \underset{C,W}{\textit{argmax}}\ P(A|W)P(W, C) \quad (2)$$

$P(A|W)$ est la probabilité estimée par les modèles acoustiques pour une chaîne de mot W. $P(W, C)$ est la probabilité jointe d'une chaîne de mots W et d'une chaîne de concepts C. Cette recherche de la meilleure interprétation $\hat{C}$ va être 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 puis en graphe de valeurs de concepts. Ce processus est présenté dans le paragraphe suivant.

3 D'UN GRAPHE DE MOTS VERS UN GRAPHE DE SOLUTIONS ENRICHI

Dans cette étude, le but est d'obtenir une séquence de concepts associés à une valeur, comme dans l'exemple présenté dans le tableau 1. Ainsi nous annoterons C l'interprétation et c_i la composante de base de cette interprétation, V la valeur associée à l'interprétation C et v_i la composante de la valeur associée à la composante de l'interprétation c_i. À chaque concept c_i correspond la chaîne de mot W^{c_i} qui contient les information nécessaires à l'extraction de la valeur v_i. Cette dernière correspond à la chaîne de mots W^{v_i}, contenue dans la chaîne de mot associée W^{c_i}. Ainsi l'interprétation I de n concept-valeur contenus dans un message est représentée à la fois par la séquence de concepts $C = \{c_1, c_2, \ldots c_n\}$, la séquence de valeur $V = \{v_1, v_2, \ldots v_n\}$ et les séquences de mots $W^C = \{W^{c_1}, W^{c_2}, \ldots, W^{c_n}\}$ et $W^V = \{W^{v_1}, W^{v_2}, \ldots, W^{v_n}\}$.

Par exemple, la chaîne W *entre Limoges* appartient à la séquence de concepts W^C associé au concept C de lieu *localisation-ville* ; la valeur V attaché à la cette même chaîne de mots W^V et appartenant au concept C de lieu est *limoges*.

Pour modeliser les c_i et les v_i nous utilisons des grammaires régulières, codées sous la forme d'automates à état fini (Finite State Machines ou FSM). c'est grâce à leur petite taille et leur simplicité que nous pouvons les représenter ainsi. Pour créer ces grammaires, deux possibilités s'offrent à nous :

– automatique : par apprentissage sur un corpus créé à cet effet, à partir du corpus MEDIA pour trouver les séquences de mots W^{c_i} associées aux c_i ; puis, nous généralisons cette chaîne (nous remplaçons certains mots comme les noms propres, les noms de villes, *etc.*) pour créer un ensemble qui va regrouper tous ces automates en un seul A_{c_i} pour les concepts, et A_{v_i} pour les valeurs associées ;
– manuellement : par des règles induite par des bases de connaissance, qui ne sont pas spécifiques au corpus MEDIA (forme des dates, des chiffres, etc.), ces automates sont utilisés pour compléter les automates obtenus par apprentissage automatique.

Nous avons montré l'intéret d'utiliser ces deux méthodes conjointement, dans l'article [Servan and Bechet, 2006].

Une fois l'ensemble des automates A_{c_i} pour les concepts et A_{v_i} pour les valeurs créés, nous les transformons en transducteurs t_{c_i} et t_{v_i} prenant en entrée comme symboles, les chaines W^{c_i} et W^{v_i}, les symboles de sortie associés à ces transducteurs sont les concepts c_i et les valeurs v_i. Il existe un transducteur particulier, celui-ci t_{bck} prend en entrée n'importe quelle chaîne de mots pour émettre le symbole de sortie BCK qui correspond au concept *background* (noté *null* dans l'onthologie MEDIA). Ce dernier permet de signaler toutes les chaines de mots qui n'ont pas été associées avec un concept. Par défaut, l'ensemble des chaines de mots peuvent être associées à ce concept. Enfin l'ensemble des transducteurs t_{c_i} sont unifiés en un transducteur unique pour les concepts, $T_{Concepts}$. De la même manière le transducteur $T_{Valeurs}$, regroupe l'ensemble des transducteurs t_{v_i} représentant le transducteur des valeurs.

Ainsi, grâce à ces deux transducteurs, nous pouvons traiter un graphe de mots G quel que soit son origine (issu de

la RAP ou issu d'un texte de transcription). En le composant avec le transducteur $T_{Concepts}$ des concepts, nous obtenons un graphe de mots enrichit des concepts C que nous appellerons $G_{Concepts}$. Ce même transducteur correspond dans le même temps à la suite de mots W^C si l'on considère les symboles d'entrée.
Le score d'un chemin dans $G_{Concept}$ est calculé à partir de l'équation 2 qui devient :

$$P(\hat{C}, \hat{W}|A) \approx \underset{W^C \in G_{Concept}}{argmax} P(A|W^C)P(W^C, C) \quad (3)$$

Le premier terme utilise les scores acoustiques contenu dans le graphe G issu de la RAP, le calcul de $P(W^C, C)$ est présenté dans la partie 4.2. Nous enumérons les n-meilleurs chemins de $G_{Concepts}$, afin d'obtenir les n-meilleures interprétations. De même, si l'on souhaite obtenir les p-meilleures chaînes de mots pour une interprétation C, il nous suffit d'énumérer les p-meilleurs chemins issus de W^C produisant C. Ce processus est détaillé dans [Raymond and Bechet, 2006].
Enfin, nous appliquons le transducteur des valeurs $T_{Valeurs}$ sur le graphe $G_{Concepts}$, pour obtenir un graphe $G_{Valeurs}$ enrichit des valeurs. Nous obtenons une graphe de mots $G_{Valeurs}$ enrichi des concepts et des valeurs associées, tout en conservant les scores du modèle acoustique et du modèle de langage. Ce transducteurs ne modifie pas les scores du graphe mais permet de faire un tri parmi les chemins présents. En effet, certaines chaines de mots peuvent être associées à des concepts mais pas à leurs valeurs. Ce constat résulte de la généralisation opérée sur les transducteurs des concepts à leur création. Ainsi, pour obtenir le graphe $G_{Valeurs}$ enrichit des valeurs, nous effectuons l'opération suivante, détaillée dans le paragraphe 5.5 :

$$G_{Valeurs} = G_{Concepts} \cap T_{Valeurs} \quad (4)$$

4 APPRENTISSAGE DES MODÈLES

4.1 Les probabilités acoustiques

Le décodeur SPEERAL [Nocera and Al., 2002] a été utilisé pour transcrire les messages du corpus MEDIA. Ces messages sont enregistrés dans des conditions identiques à celle que l'on peut trouver dans un système mis en service. Les utilisateurs ont effectués leurs appels depuis leur téléphone, fixe ou cellulaire, et la qualité des enregistrements est variable. Les modèles acoustiques téléphoniques utilisés sont ceux développées 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 (Maximum A Posteriori).
Le modèle de langage 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 est de 32,2%.

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

4.2 Le modèle de langage

La probabilité de la chaine de mots W^C et de la séquence de concepts C, permet de calculer conjointement la probabilité $P(W^C, C)$. Nous pouvons maintenant mettre une étiquette t_i sur chaque mot w_i. Si $w_i \in c_j$, alors $t_i = c_j$. Dans le cas contraire, w_i n'appartient à aucun concept et se retrouve dont avec l'étiquette $t_i = null$. Ainsi, nous aurons :
$P(W^C, C) = P\{(t_1, w_1)(t_2, w_2)\dots(t_n, w_n)\} = P(t_{1,n}, w_{1,n})$.
Ce processus est identique à la problématique des étiqueteurs probabilistes, telle qu'on peut la trouver dans [Charniak and Al., 1993]. En définissant de manière adéquate des termes tels que $t_{1,0}$, ainsi que leurs probabilités, on obtient :

$$P(t_{1,n}, w_{1,n}) = \prod_{i=1}^{n} P(t_i|t_{1,i-1}, w_{1,i-1})P(w_i|t_{1,i}, w_{1,i-1}) \quad (5)$$

De manière à pouvoir estimer ces probabilités, nous faisons les hypothèses de Markov suivantes :

$$P(t_i|t_{1,i-1}, w_{1,i-1}) = P(t_i|t_{i-2,i-1}, w_{i-2,i-1}) \quad et \quad P(w_i|t_{1,i}, w_{1,i-1}) = P(w_i|t_{i-2,i}, w_{i-2,i-1}) \quad (6)$$

Ainsi suivant les propriétés des modèles de Markov cachés (HMM) trigrammes, t_i ne dépend plus que des deux étiquettes précédentes tout comme w_i dépend des deux mots précédents
Nous obtenons l'équation suivante :

$$P(t_{1,n}, w_{1,n}) = \prod_{i=1}^{n} \{P(t_i|t_{i-2,i-1}, w_{i-2,i-1}) \cdot P(w_i|t_{i-2,i}, w_{i-2,i-1})\} \quad (7)$$

Afin d'apprendre les probabilités de l'équation 7, un corpus d'apprentissage contenant des transcriptions de dialogues est nécessaire. Ce corpus doit être manuellement étiqueté en concepts, il est ensuite formaté comme dans l'exemple précédent. Les probabilités de l'équation 7 sont alors apprises selon le critère du maximum de vraisemblance, avec un nécessaire lissage pour les n-grammes non vus. On obtient de cette manière un modèle de langage avec repli qui est utilisé pour estimer la probabilité $P(W^C, C)$ de l'équation 3. Ce modèle de langage est représenté sous forme d'automates grâce à l'ensemble d'outils *FSM/GRM library* d'AT&T [Mohri and Al., 2002], il est composé avec le transducteur $G_{Concept}$ pour donner le score final à chaque chemin de $G_{Concept}$.

5 DESCRIPTION DU SYSTÈME

5.1 L'apprentissage

La partie apprentissage s'appuit comme présentée dans le paragraphe 2.4 sur des HMM et plus particulièrement une des modélisations de ces modèles sur les FSM. Pour chaque transducteur $T_{Concepts}$ et $T_{Valeurs}$ nous avons un corpus spécifique, issu du corpus d'apprentissage MEDIA. Ces corpus sont mis aux formats de décodage voulus respectivement pour les *concepts* et les *valeurs*. Le modèle de langage dont la construction est expliquée dans le paragraphe 4.2, est lui aussi appris sur un corpus spécifique, issu du corpus MEDIA, pour la création d'un modèle de langage trigramme.

5.2 Format des étiquettes

Le formalisme des étiquettes t_i est le suivant :

– c_i_I : qui indique que le mot étiqueté appartient au concept c_i (I pour inside, "dans" en anglais) ;
– c_i_B : le mot étiqueté ainsi se trouve en début de concept (B pour begin, "début" en anglais).

Par exemple la séquence *euh le deux mars à midi* pourra être étiquetée :

```
(euh,null)(le,null)(deux,$TMP_DTE_B)(mars,TMP_DTE_I)
(à,null)(midi,$TMP_PTM_B)
```

Afin de limiter le phénomène du manque de données, certains mots sont généralisés grâce à un ensemble de symboles non-terminaux correspondant aux chiffres, aux jours de la semaine, aux mois, *etc*. Ainsi l'exemple précédent devient :

```
(euh,null)(le,null)($NB,$TMP_DTE_B)($MONTH,$TMP_DTE_I)
(à,null)(midi,$TMP_PTM_B)
```

5.3 Les règles à base de connaissance

Chaque transducteur $T_{Concepts}$ et $T_{Valeurs}$ est composé d'une grande partie d'apprentissage mais aussi d'une partie formelle. En effet, en utilisant la librairie FSM/GRM [Mohri and Al., 2002] nous pouvons construire des grammaires formelles pour créer des transducteurs de concepts t_{c_i} et de valeurs t_{v_i}. Ces données *a priori* sont en fait des règles de grammaire qui ne sont pas spécifiques au corpus MEDIA, ici nous les avons utilisées principalement pour la représentation des dates et des nombres, ainsi que pour l'attibution des valeurs associées à ces concepts. Enfin, nous unissons les transducteurs de grammaires formelles de concepts et de valeurs respectivement aux transducteurs $T_{Concepts}$ et $T_{Valeurs}$ les transducteurs de concepts et de valeurs.

5.4 Le décodage

Le décodage montré schématiquement dans la figure 5.4 a été décomposé pour simplifier la compréhension. Nous pouvons voir sur cette figure, la décomposition de l'opération de composition de l'ensemble de tous les transducteurs sur le graphe de mots G. Comme dit précédemment, celui-ci peut être issu soit d'un module de RAP, soit des transcriptions. Nous appliquons sur ce graphe successivement le transducteur de concepts $T_{Concepts}$, pour obtenir $G_{Concepts}$ puis le modèle de langage pour appliquer les probabilités du modèle de langage, et enfin le transducteur des valeurs $T_{Valeurs}$ pour passer au graphe de valeurs $G_{Valeurs}$. À l'issue de ces compositions, nous pouvons déterminer la ou les meilleures interprétations conceptuelles (notées *N-Best*) ; les transducteurs $T_{Concepts}$, le modèle de langage et $T_{Valeurs}$ sont notés respectivements *FSM Concepts FSM Taggeur* et *FSM Valeurs* dans la figure 5.4

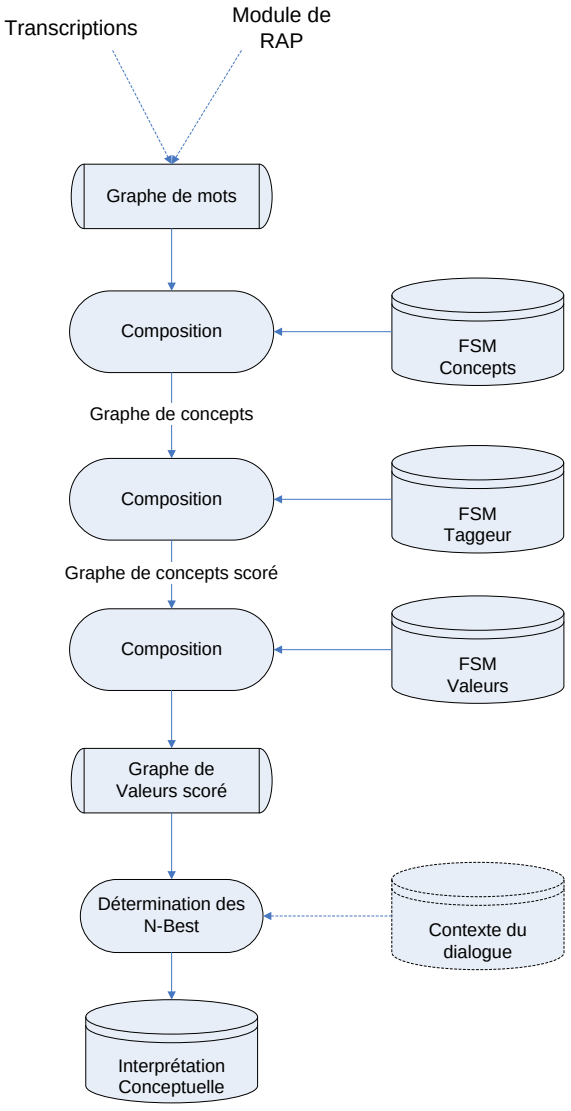

FIG. 1 – Stratégie de décodage

5.5 d'un graphe de concepts à un graphe de valeurs

Les valeurs v_k sont une extension des étiquettes attribuées t_i à chaque mot w_i. Chaque v_k dépend du concepts c_j associé à w_i. De ce fait, le graphe de valeur $G_{Valeurs}$ n'existe que si pour un mot w_i donné, celui-ci appartient à c_j et s'il appartient à v_k et si v_k appartient à c_j :
$il\ existe\ G_{Valeurs}\ tel\ qu'il\ existe\ un\ v_k \in w_i$
$tel\ que\ v_k \in c_j\ existe.$
Dans le cas contraire, comme pour l'attribution des

étiquettes, si w_i n'a pas de valeur correspondant à son concepts, w_i se retrouve avec l'étiquette $t_i = null$ et avec aucune valeur associée.

Ce passage aux valeurs se fait grâce au transducteur des valeurs $T_{Valeurs}$. Nous l'appliquons au graphe de concepts $G_{Concepts}$ pour pouvoir ainsi filtrer les symboles d'entrée possible dans les valeurs. Nous avons comme symbole d'entrée, les mots associés à leur étiquettes, et en symboles de sortie, les valeurs associées à ces composition de mots et d'étiquettes. Nous obtenons à partir de la transduction de $G_{Concepts}$, par $T_{Valeurs}$ un graphe de sortie $G_{Valeurs}$ qui a conservé les poids de $G_{Concepts}$. Ainsi nous pouvons obtenir les n-meilleures interprétations contenues dans le graphe $G_{Valeurs}$, mais aussi, grâce à ses symboles d'entrées, nous pouvons aussi obtenir la chaine de mots W^V associée à la meilleure interprétation conceptuelle avec les valeurs V.

5.6 Format de sortie du graphe de valeurs

L'ensemble des valeurs étant particulièrement grand, nous avons décidé de procéder par identification, pour limiter la taille des lexiques associés aux transducteurs. À chaque valeur est associé un identifiant unique, par exemple :

```
997 cardinal
998 carlton
999 carmes
1000 carpentras
```

Tout ce qui est traitement des nombres (tarifs, nombre de chambres) ainsi que les dates, ce sont les valeurs qui sont directement associées aux étiquettes et aux mots.

Ainsi en reprenant l'exemple *euh le deux mars à midi* qui avait été étiqueté :

```
(euh,null)(le,null)(deux,$TMP_DTE_B)(mars,$TMP_DTE_I)
(à,$LOC_VIL_B)(carpentras,$LOC_VIL_I)
```

Avec l'application du transducteur des valeurs devient :

```
(euh,null)(le,null)(deux,$TMP_DTE_B#02)
(mars,$TMP_DTE_I#03)(à,$LOC_VIL_B#1000)
(carpentras,$LOC_VIL_I#1000)
```

le numéro "1000" correspondant à la valeur de "carpentras", les autres concepts étant des concepts associés à des nombres que nous conservons tels quels dans notre décodage, jusqu'au passage au format MEDIA.

5.7 le format MEDIA

Le format de sortie MEDIA est une simple mise en forme du tableau 1. Nous utilisons un format assimilé XML, défini par l'article [MEDIA, 2005] pour obtenir une plus grande lisibilité de la sortie du système, précedemment montrée. Ainsi toujours en conservant l'exemple *euh le deux mars à midi*, voici ce que devient cette phrase après la sortie système :

```
<Turn uid="1234" startTime="2.492" endTime="9.726"
speaker="spk5">
<SemDebut identifiant="1" mode="+" concept="null"
valeur="1" reference=""/>
euh le
<SemFin/>
<SemDebut identifiant="2" mode="+" concept="temps
-date" valeur="02/03" reference=""/>
deux mars
<SemFin/>
<SemDebut identifiant="3" mode="+" concept="locali-
sation-ville" valeur="carpentras" reference=""/>
 à Carpentras
<SemFin/>
</Turn>
```

6 LES EXPÉRIENCES MENÉES

Les expériences ont été menées sur le test hors-contexte issu de la campagne EVALDA/MEDIA [MEDIA, 2006]. Nous utilisons les 83 attributs, présentés au paragraphe 2.2 pour l'évaluation de notre système. Dans cette étude, nous ne tiendrons pas compte des 19 spécifieurs ainsi que du mode. L'attribution des spécifieurs ainsi que du mode est faire par un module après le choix de l'interprétation. Ils ne sont pas concernés par le décodage conceptuel présenté ici. Le corpus d'apprentissage est découpé en 720 dialogues tous annotés,

Ces expériences ont été faites sur les transcriptions d'une part puis sur la sortie du module RAP (paragraphe 4.1) d'autre part.

6.1 Le *Concept Error Rate*

Pour mesurer les performance, nous utilisons le taux d'erreur sur les paires attributs-valeurs, que l'on appelle le *Concept Error Rate* ou *CER*. Une interprétation est considérée comme correcte si et seulement si l'attribut et la valeur mis au format MEDIA sont corrects par rapport à la référence.

En alignant l'interprétation de référence et celle obtenue par notre décodage, nous pouvons calculer le nombre de paires concept-attribut correctes C ainsi que le nombre des différents types d'erreurs :

- insertion I, lorsqu'un concept est inséré par erreur ;
- deletion D, si un concept est oublié ;
- substitution S, tant pour l'attribut que pour la valeur.

Avec R étant le nombre de concepts contenus dans la référence, nous pouvons calculer le CER avec la formule (similaire au *WER*, le *word error rate*) :

$$CER = \frac{I + D + S}{R} \times 100$$

.

Nous pouvons voir dans le tableau 2, les résultats comparatifs entre la sortie issue de la RAP et les transcriptions. Nous avons comparé dans un premier temps la meilleure sortie du système de RAP, et les transcriptions. Nous avons constaté que notre système est performant sur les transcriptions manuelles, celui ci donnant des résultats identiques aux meilleurs systèmes présents lors de l'évaluation MEDIA. Nous pouvons voir que le passage à la parole dégrade les résultats, le taux d'erreur

Graphes	$G_{1-BestRAP}$	$G_{GrapheRAP}$	G_{Trans}
WER	35.5	32,2	19.8
WER Oracle de taille 20	28.5	20.5	10.2

TAB. 2 – Résultats obtenus sur la partie test du corpus MEDIA

mot sur la meilleure sortie du module de RAP étant de 32.2%.
Aux vues de ces résultats nous avons décidé d'utiliser le graphe complet de la sortie du module de RAP. Nous avons ainsi pu constater une amélioration des résultats.

6.2 Le taux oracle

Suite à cette expérience, nous avons décidé d'exploiter le graphe complet en observant la liste des n-meilleures interprétations différentes produites. Cette liste est particulièrement intéressante lorsqu'on se place dans le contexte d'un dialogue avec une machine. Ainsi, il est possible de fournir au gestionnaire de dialogue un ensemble d'hypothèses. Ce dernier pourra les filtrer grâce à un éventuel contexte de dialogue. La mesure communément utilisée pour juger de la qualité d'un graphe ou d'un ensemble d'hypothèses est la mesure du taux *oracle*. Le principe de cette mesure est de choisir parmi toutes les hypothèses proposées, celle qui a le plus petit taux d'erreur. Le résultat donne le taux d'erreur minimal que ferait le système s'il choisissait toujours la bonne solution parmi la liste proposée.
Nous avons réalisé une liste d'hypothèses pour le graphe de parole $G_{GrapheRAP}$ ainsi que pour la meilleure sortie du module de RAP $G_{1-BestRAP}$ que nous avons comparé aux transcripions G_{Trans}. Les résultats de l'évaluation sont présentés dans le tableau 2 et illustrés avec la figure 6.2.

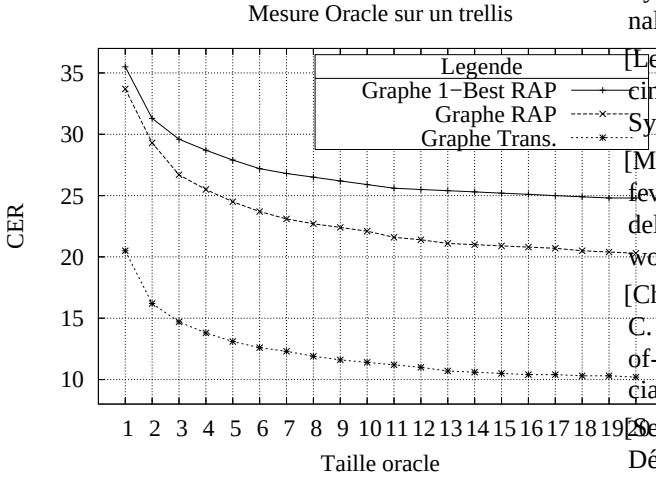

FIG. 2 – CER en variant l'oracle de 1 à 20

7 CONCLUSION ET PERSPECTIVES

Nous avons présenté dans cette étude, un système de décodage conceptuel basé sur une approche stochastique composé uniquement d'automates à états finis (*FSM* ou *Finite State Machines*). Le principal intéret d'utiliser ces FSM est de conserver le graphe de solutions issu de la RAP. En conservant, l'espace complet après avoir appliqué toutes les transductions, nous nous réservons le choix de ou des meilleures interprétations à un niveau de décision pouvant intégrer un contexte de dialogue.
Les expériences menée sur le corpus MEDIA, permettent de mettre en évidence qu'un décodage conceptuel utilisant exclusivement des transducteurs donne des résultats très intéressants, de l'ordre de ceux présents lors de l'évaluation MEDIA [MEDIA, 2006]. Nous avons montré qu'un décodage cherchant la meilleure interprétation dans un graphe complet de valeurs et de concepts lui-même issu d'un graphe de complet de solutions créé par le module de RAP est plus efficace d'après les résultats obtenus, qu'un décodage appliqué sur la meilleure solution issu du module de RAP. Ainsi ce graphe de mots contenant une plus grande diversité d'interprétations, a plus de chance de contenir l'interprétation correcte que nous recherchons.
Enfin, la création d'une liste *oracle* dans le cadre d'un décodage conceptuel, donne la possibilité d'intégrer celle-ci dans un système de dialogue. En effet, en tenant compte conjointement d'une liste de meilleures interprétation et du contexte de dialogue associé à cette liste, nous pourrions améliorer la prise de décision.

BIBLIOGRAPHIE

[Denis and Al., 2006] Denis A. and Quinard M. and Pitel G., A Deep-Parsing Approach to Natural Language Understanding in Dialogue System : Results of a Corpus-Based Evaluation, LREC 2006, Genoa Italy.

[Seneff, 1992] Seneff S., TINA : A Natural Language System for Spoken Language Applications, Computational Linguistics, 18, 61-86.

[Levin and Al., 1995] Esther Levin and Roberto Pieraccini, Concept-Based Spontaneous Speech Understanding System, EUROSPEECH 1995, 555-558, Madrid Spain.

[Maynard and Al., 2005] Bonneau-Maynard H. and Lefevre F., A 2+1-Level Stochastic Understanding Model, Automatic Speech Recognition and Understanding workshop (ASRU) 2005.

[Charniak and Al., 1993] Charniak E. and Hendrickson C. and Jacobson N. and Perkowitz M, Equations for Part-of-Speech Tagging 11th National Conference on Artificial Intelligence 1993, 784-789

[Servan and Bechet, 2006] Servan C. and Bechet F., Décodage conceptuel et apprentissage automatique : application au corpus de dialogue Homme-Machine MEDIA, TALN 2006, Leuven Belgium.

[MEDIA, 2005] Bonneau-Maynard H. and Rosset S. and Ayache C. and Kuhn A. and Mostefa D., Semantic annotation of the French Media dialog corpus, Proceedings

of the European Conference on Speech Communication and Technology (Eurospeech 2005), Lisboa Portugal.

[MEDIA, 2006] Bonneau-Maynard H. and Ayache C. and Béchet F. and Denis A. and Khun A. and Lefevre F. and Mostefa D. and Quinard M. and Servan C. and Villaneau J., Results of French Evalda-Media evaluation campaign for litteral understanding, LREC 2006, 2054-2059.

[Raymond and Bechet, 2006] Raymond, C. and Béchet F. and De Mori R. and Damnati G., On the use of finite state transducers for semantic interpretation, Speech Communication 2006, 48,3-4, 288-304.

[Mohri and Al., 2002] Mohri M.and Pereira F. and Riley M., Weighted Finite-State Transducers in Speech Recognition, Computer Speech and Language, 16, 69-88.

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