Composition sémantique pour la compréhension de la parole dans un cadre de dialogue

Frédéric Duvert *Marie-Jean Meurs* *Christophe Servan* *Frédéric Béchet*
Fabrice Lefèvre *Renato de Mori**

Laboratoire d'Informatique d'Avignon
339, chemin des Meinajaries
Agroparc BP 1228
84911 AVIGNON Cedex 9
FRANCE

ABSTRACT

A knowledge representation formalism for SLU is introduced. It is used for incremental and partially automated annotation of the MEDIA corpus in terms of semantic structures. An automatic interpretation process is described for composing semantic structures from basic semantic constituents using patterns involving constituents and words. The process has procedures for obtaining semantic compositions and for generating Frame hypotheses by inference. This process is evaluated on a dialogue corpus manually annotated at the word and semantic constituent levels.
Keywords : **Spoken language understanding, semantic structures, Frames, conceptual decoding, semantic annotation, semantic inference.**

1. Introduction

La compréhension du langage parlé regroupe l'ensemble des analyses visant à caractériser, étiqueter, structurer et finalement représenter formellement l'information contenue dans un message vocal. Les relations sont représentées par une Source de Connaissance (SC) et appliquées par des processus utilisant des stratégies de contrôles de ces mêmes connaissances. Cette tâche est ardue car le processus de compréhension est imbriqué à l'intérieur d'une chaîne de traitement regroupant plusieurs tâches telles que le traitement de signal et la Reconnaissance Automatique de la Parole (RAP), chacune pouvant générer ses propres erreurs dues à des problèmes de modélisation ou de manque de robustesse au bruit (bruit acoustique ou lexical).

De plus le langage naturel spontané qui est la cible de cette étude est caractérisé par des phrases difficiles à traiter, souvent sans forme grammaticale valide. Les hypothèses de mots des transcriptions automatiques produites par les systèmes de RAP contiennent aussi des erreurs, rendant difficile tout traitement linguistique.

Dans le but de minimiser les effets de ces imprécisions, l'interprétation doit être conçue comme un processus de décision qui peut être décomposé conceptuellement en sous-tâches. Il a été observé qu'une augmentation de la précision peut être atteinte en considérant un treillis d'hypothèses de constituants sémantiques à partir d'un treillis d'hypothèses de mots [7]. Les constituants sémantiques sont ensuite *composés* à l'intérieur de structures sémantiques de plus haut-niveau. Les hypothèses sur les composants sémantiques sont générés en utilisant des machines à états finis, suivant les travaux de [8].

Ce papier décrit une nouvelle combinaison de constituants sémantiques dans des structures sémantiques ainsi qu'un processus d'évaluation. Les constituants sont générés par un processus de traduction à partir d'un treillis de mots. Ces constituants sont ensuite caractérisés par des spécifieurs de sens s'étendant à tout le message. Les structures sémantiques de haut niveau sont obtenues à partir de règles s'appliquant sur ces constituants et ces spécifieurs. Cette approche a été testée sur un corpus français de dialogue oral, le corpus MEDIA. Les résultats expérimentaux issus de cette approche sont présentés en fin de papier.

2. Le corpus Media

Le corpus MEDIA [2] a été enregistré en utilisant un système de simulation (*Magicien d'Oz*) de serveur vocal téléphonique pour des informations touristiques et réservations d'hôtels. Huit catégories de scénario ont été définies avec des degrés de complexités différents. Le corpus compte 1257 dialogues, 250 interlocuteurs et contient environ 70 heures de dialogues. La portion d'entraînement du corpus est riche de plus de 80 concepts de base transcrits et annotés manuellement.

Cette représentation sémantique *plate* est enrichie par des étiquettes qui peuvent être considérées comme des traces de la représentation hiérarchique sous-jacente. La représentation hiérarchique permet de représenter explicitement des relations entre segments potentiellement non-adjacents dans la transcription d'une requête. Cependant une représentation *plate* facilite l'annotation manuelle des données. Il a été décidé pour le schéma d'annotation de MEDIA de préserver les relations en définissant des spécifieurs qui sont combinés avec les rôles de bases. Il y a 19 spécifieurs dans le modèle sémantique de MEDIA.

Un exemple de l'annotation de MEDIA est donné dans le tableau 1. Le spécifieur *réservation* est attribué aux concepts *commande*,*nombre-chambre*, *nombre-nuit* comme structure hiérarchique. Cela représente une réservation déclenchée par le concept *commande* et remplie avec les éléments trouvés dans

*Ces travaux sont supportés par le 6ième Programme de Recherche de l'Union Européenne, Projet LUNA, IST,contrat: 33549. Pour plus d'informations sur le Projet LUNA, veuillez visiter [1].

Tab. 1: Exemple d'annotation sémantique du corpus MEDIA

n	W^{c_n}	c_n	*spécifieur*	*valeur*
1	*he bien*	null		
2	*je souhaiterais réserver*	commande		réservation
3	*à l'hôtel Richard Lenoir*	nom-hotel		richard_lenoir
4	*six*	nombre-chambre	réservation	6
5	*chambres individuelles*	type-chambre		simple
6	*à partir du trente-et-un mai*	date	réservation	31/05
7	*deux jours hum deux nuits*	nombre-nuit	réservation	2

commande,*nombre-chambre* et *nombre-nuit*.

La combinaison des spécifieurs et des attributs permet de recomposer une structure hiérarchique d'une requête à partir de sa représentation plate. Cette annotation fournit des étiquettes comparables aux constituants sémantiques hypothétisés par un parseur sémantique de surface.

La combinaison des rôles de base et des spécifieurs donne 1121 attributs potentiels. Un total de 144 attributs distincts apparaît dans le corpus d'entraînement avec environ 2200 valeurs normalisées différentes.

3. Décodage conceptuel pour la génération de constituants basiques

Le corpus MEDIA est annoté avec des composants sémantiques basiques mais pas avec des structures sémantiques. Les constituants sémantiques basiques sont hypothétisés et testés suivant l'approche décrite dans [7].

Le processus de décodage conceptuel est vu comme un processus de transduction dans lequel les modèles de langages stochastiques sont implémentés par des automates à états finis (*Finite State Machines* ou FSM). Ils produisent des étiquettes pour les composants sémantiques. Il y a un FSM pour chaque composant conceptuel élémentaire. Chaque FSM implémente une grammaire régulière modélisant les différentes formes supports pour chaque composant conceptuel. Ces FSMs sont des transducteurs qui prennent les mots en entrée et produisent des étiquettes conceptuelles. Ils sont appliqués à un graphe de mots sorti par le module de RAP grâce à une opération de composition entre automate.

Pour trouver la meilleure séquence de concepts et de mots, un étiqueteur statistique à base de HMM (lui aussi représenté par un FSM) est utilisé pour réévaluer chaque chemin dans le graphe de mots/concepts. Cet étiqueteur est entraîné sur le corpus d'apprentissage de MEDIA. Cette approche est appelée *décodage intégrée*, puisque les processus de RAP et de compréhension du langage sont réalisés en même temps en recherchant de manière simultanée la meilleure séquence de mots et de concepts. Le résultat du processus de traduction est une liste structurée des n-meilleures interprétations qui peuvent être vues comme toutes les interprétations possibles d'un énoncé.

4. Spécification du sens

Les interprétations conceptuelles issues de la liste des n-meilleures hypothèses n'ont pas d'étiquettes de spécification du sens. Ces spécifieurs sont ajoutés dans une seconde phase par un processus de marquage basé sur des classifieurs discriminants [5]. Les *Conditional Random Fields* (CRF) [4] retenus dans notre étude, ont été largement utilisés pour différents traitements d'étiquetage de mots comme les *Part-Of-Speech* ou pour la détection des entités nommées. L'utilisation des CRFs est une approche discriminante qui a donné de meilleurs résultats sur ce genre de tâches que des approches génératives basées sur des HMMs. Le principal avantage des CRFs est la capacité de prédire une étiquette en tenant compte de l'ensemble du message. C'est très important pour l'ajout de spécifieurs aux concepts, car cette information dépend des éléments qui peuvent être très éloignés du concept à étiqueter dans le message.

L'entraînement pour l'étiquetage des spécifieurs se fait sur le corpus MEDIA, chaque message est une séquence d'éléments (mots, attributs, valeurs) étiquetée par un spécifieur ou par le symbole *NULL*. Lors du décodage, chaque séquence hypothèse de mots/concepts de la liste de n-meilleures séquences est traitée par l'étiqueteur pour ajouter les spécifieurs. L'outil **CRF++** [1] est utilisé.

5. Représentation sémantique structurée

Les structures sémantiques peuvent être dérivées à partir d'une source de connaissance obtenue par une théorie sémantique. Par exemple les réseaux sémantiques qui représentent les entités et leurs relations [9] ou des structures fonction/arguments [3]. Un moyen pratique de représenter et raisonner sur les sources de connaissances est l'utilisation d'ensemble de *formules logiques*. On peut ainsi les dériver en des structures informatiques telles les Frames. Une Frame est un modèle pour la représentation sémantique d'entités et leurs propriétés. Les Frames sont capables de représenter des types des structures conceptuelles aussi bien que des instances de ceux-ci.

Un élément de Frame est une structure de données de la Frame. Il décrit :
– les propriétés d'une structure sémantique,
– les valeurs (qui peuvent être d'autres instances de Frame),

[1] http ://crfpp.sourceforge.net/

– les contraintes qui doivent être respectées par les valeurs,
– les procédures pour obtenir les valeurs des propriétés à partir des observations.

L'obtention des éléments d'une Frame engendre l'instanciation de cette Frame. Une grammaire de Frame permet de caractériser l'ensemble des Frames acceptable pour la représentation sémantique d'un domaine.

6. Annotation progressive du corpus par des structures sémantiques

Une ontologie ou source de connaissance basée sur les Frames a été écrite manuellement pour décrire le domaine sémantique du corpus MEDIA. Quelques Frames décrivent des connaissances génériques comme des relations spatiales, d'autres sont spécifiques à l'application. Ces Frames sont définies en tenant compte du paradigme *FrameNet* de Berkeley adopté par [1].

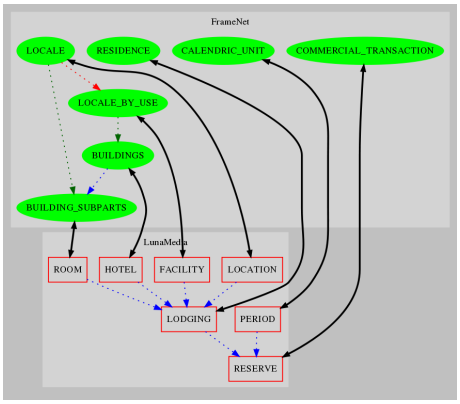

Fig. 1: Représentation en Frame d'une partie de la base de connaissance et sa projection de *FrameNet* vers MEDIA

La figure 1 montre un exemple d'une représentation sémantique dans le corpus MEDIA, en partant de la représentation FrameNet.

L'ontologie MEDIA est composée de 21 Frames de bases avec un total de 85 éléments de Frames. La représentation du langage contient des composants conceptuels et des procédures de compositions de structures sémantiques. Les règles de composition peuvent prendre en entrée les étiquettes des constituants conceptuels et les spécifieurs de sens, ainsi que les Frames déjà instanciées. On s'attache à décrire les éléments du message par des Frames et de les relier entre eux, une signfication du sens plus poussée des éléments de MEDIA et la modification de son ontologie sont en cours, afin d'obtenir des annotations de qualité supérieures.

Quand une règle s'applique, des instances de Frames sont créées. Des inférences basées sur les instances de Frames sont alors réalisées. Différentes Frames reliées par des relations peuvent être instanciées pour un seul message. 463 tours de dialogue issus de 15 dialogues ont été annotés manuellement. Le formalisme d'annotation de *FrameNet* [6] a été utilisé. Un outil de visualisation de Frames appelé *FriZ*, dédié au traitement du langage dans le dialogue a été développé comme support de l'annotation manuelle et de la vérification pour d'ultérieures annotations automatiques. Le temps moyen d'annotation manuelle par dialogue est d'environ 2 heures.

Par exemple la phrase "*J'accepte la réservation*" est annotée avec trois Frames :

```
ACCEPTER[(est_un:verbe)(sujet:person)(theme:reservation)]
PERSONNE [(est_un:humain) (categorie:utilisateur)...]
RESERVATION [(est_un:objet) ...]
```

Les règles sont généralisées par l'annotation progressive des données avec les connaissances disponibles. On évalue la confiance des résultats et on annote des exemples manuellement si la confiance est trop faible.

Les procédures attachées sont intégrées dans le processus d'interprétation pour fournir automatiquement des annotations de Frames sur le corpus d'entraînement ainsi que des instances d'hypothèses sur le corpus de test. Le processus est capable de réaliser des inférences sur des Frames dont l'instance est sous-entendue par une autre instance de Frame. Une centaine de règles génère les instances à partir de combinaisons de mots ou de motifs de constituants sémantiques et réalise les inférences sur les résultats. Il y a environ 30 formules d'inférences utilisées dans le processus.

Lors du décodage, une fois la liste des n-meilleures interprétations obtenue avec les spécifieurs présentés en Section 2, chaque séquence de mots/concepts est analysée par les règles produites pour le corpus d'entraînement de MEDIA. Ces règles utilisent les attributs, les valeurs et les spécifieurs obtenus lors de la première phase du décodage pour inférer des Frames.

7. Résultats expérimentaux

Les tests ont été réalisés sur un corpus de 1249 tours de dialogues pour un total de 2938 composants sémantiques. Le tableau 2 donne le taux d'erreurs obtenu après la phase de décodage conceptuel. Pour un taux d'erreurs sur les mots de 30.3%, le taux d'erreurs sur les composants conceptuels (ou *concepts*) est d'environ 25%. Chaque information, telle que les spécifieurs et les valeurs normalisées, ajoute approximativement 6% aux taux d'erreurs. Les taux d'erreurs *Oracle* obtenus en sélectionnant manuellement les meilleures hypothèses issues de la liste des n-meilleures interprétations (avec $n = 20$) sont inférieurs d'une valeur absolue de 8% par rapport au taux d'erreurs de la 1-meilleure interprétation.

Les hypothèses de Frames obtenues sur la sortie du processus d'interprétation ont aussi été évaluées. Les annotations manuelles en Frames n'étaient pas disponibles pour le corpus de test. Les annotations manuelles des mots et des concepts ont été utilisées pour dériver des annotations références en Frames. La composition et l'inférence des connaissances, décrites dans la section précédente, ont été appliquées.

Un échantillonnage aléatoire sur des tours de dia-

	corr(%)	sub(%)	suppr(%)	ins(%)	ER(%)	OER(%)
mot	75.9	15.3	8.8	6.2	30.3	22.5
concept	85.0	8.7	6.3	10.3	25.3	19.2
+ specif	78.6	15.2	6.2	10.2	31.6	23.4
+ value	72.5	21.4	6.1	10.1	37.6	25.2

Tab. 2: Taux d'erreurs (ER) et Taux d'erreurs Oracle (OER) sur la n-meilleure liste des interprétations pour les mots, les concepts, les concepts avec des spécifieurs et valeurs

logues utilisateurs a été prélevé pour estimer, par deux experts humains, l'exactitude de l'annotation automatique des structures sémantiques. Une F-mesure de 0.90 (précision de 0.96, rappel de 0.85) a été mesurée sur 100 tours de dialogues, par comparaison avec les annotations manuelles et l'annotation automatique en Frame sur des transcriptions exactes. Cette forte précision permet d'utiliser une annotation automatique comme annotation de référence.

La composition et l'inférence ont été appliquées automatiquement sur la liste des n-meilleures interprétations. L'évaluation a été faite en estimant la précision, le rappel et la F-mesure sur la détection des types corrects de Frames en utilisant les annotations automatiques de références décrites plus haut. La F-mesure Oracle de la liste des n-meilleures interprétations est donné par la figure 2.

Une F-mesure de 0.92 (précision de 0.90, rappel de 0.94) a été obtenue sur la meilleure hypothèse d'interprétation, pour les 1249 tours de dialogues. Ces résultats tendent à montrer que les annotations sémantiques de haut-niveau (identifiant des Frames) sont robustes aux erreurs de RAP, les erreurs d'interprétations apparaissant le plus souvent au niveau des éléments des Frames. Sur un même processus d'annotations de la réference et des n-meilleures interprétations, il y a peu d'erreurs sur les n-meilleures interprétations issus du décodage conceptuel (tableau 2).

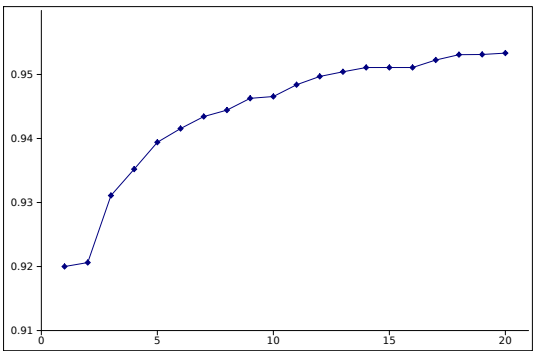

Fig. 2: F-mesure Oracle pour l'identification des Frames instanciées sur les premières n-meilleures séquences d'éléments conceptuels extraits à partir d'un treillis de décodage, en fonction de n.

La prochaine étape de notre étude sera d'exploiter la liste des n-meilleures interprétations dans le but de corriger les erreurs sur les éléments de Frames en considérant le contexte du dialogue.

8. Conclusion

Un formalisme de représentation des connaissances pour la compréhension de la parole a été introduit. Il a été utilisé pour une annotation incrémentale et automatique du corpus MEDIA en termes des structures sémantiques. Les annotations automatiques ont été évaluées et soumises à des experts humains quand la confiance était faible. Un processus d'annotation automatique a été introduit pour composer des structures sémantiques à partir d'éléments sémantiques basiques, en utilisant des règles impliquant des constituants conceptuels et des spécifieurs de sens. Le traitement utilise des procédures pour obtenir des compositions sémantiques et générer les hypothèses de Frames par inférences.

Les résultats présentés montrent que la source de connaissance et les procédures attachés ont de bonnes capacités à produire des hypothèses de structures sémantiques. Ces travaux de recherche seront poursuivis en utilisant des structures sémantiques pour sélectionner les possibles composants sémantiques au-delà de la 1-meilleure interprétation dans tout le treillis d'hypothèse de concepts.

Références

[1] Project luna : www.ist-luna.eu.

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

[3] R. Jackendoff. Semantic structures. *The MIT Press, Cambridge Mass.*, 1990.

[4] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[5] Fabrice Lefèvre. Dynamic bayesian networks and discriminative classifiers for multi-stage semantic interpretation. In *ICASSP*, Hawaii, USA, 2007.

[6] J.B. Lowe, C.F. Baker, and C.J. Fillmore. A frame-semantic approach to semantic annotation. In *Proceedings of the SIGLEX Workshop on Tagging Text with Lexical Semantics : Why, What, and How ?*, Washington D.C., USA, April 1997.

[7] 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.

[8] Giuseppe Riccardi and Al Gorin. Stochastic language adaptation over time and state in natural spoken dialogue systems. *IEEE Trans. on Speech and Audio Processing*, 8(1) :3–10, 2000.

[9] W.A. Woods. *What's in a Link : Foundations for Semantic Networks.* Bolt, Beranek and Newman, 1975.