Interprétation du dialogue oral : pour une approche bayésienne de la composition sémantique.

Marie-Jean Meurs — Fabrice Lefèvre — Renato de Mori

Université d'Avignon et des Pays de Vaucluse
Laboratoire Informatique d'Avignon (EA 931), F-84911 Avignon, France.
{marie-jean.meurs, fabrice.lefevre, renato.demori}@univ-avignon.fr

RÉSUMÉ. *Cet article présente un processus d'interprétation stochastique pour la composition de structures sémantiques, dédié aux dialogues oraux homme-machine. Il produit ces "cadres sémantiques" (CS) à partir des séquences de mots et de concepts représentant les propos du locuteur. Dans un but d'évaluation, un système à base de règles fournit une annotation de référence en CS sur le corpus d'entraînement. Grâce à un décodage utilisant des réseaux bayésiens dynamiques (DBN), le processus présenté produit des CS sur les données de test dans des conditions de difficulté croissante, fonction des erreurs commises par les systèmes automatiques de transcription et d'annotation. Les résultats expérimentaux montrent que l'approche probabiliste proposée fournit une annotation en CS de fiabilité comparable à celle d'une approche à base de règles.*

ABSTRACT. *This paper introduces a stochastic interpretation process for composing semantic structures, dedicated to spoken language interpretation. It allows to derive semantic frame structures from word and basic concept sequences representing the users' utterances. First, a rule-based process provides a reference semantic frame annotation of the training data. Then, through a decoding stage, dynamic Bayesian networks hypothesize frames from test data. Tests are performed under different conditions raising in difficulty wrt the errors due to ASR and SLU modules. The experiment results show that the proposed probabilistic framework carry out semantic frame annotation with a good reliability, comparable to a rule-based approach.*

MOTS-CLÉS : *système de dialogue oral, cadres sémantiques, composition sémantique, réseaux bayésiens dynamiques.*

KEYWORDS: *spoken dialog system, semantic frames, semantic composition, dynamic Bayesian networks.*

1. Introduction

Les méthodes stochastiques pour la compréhension automatique de la parole (*Spoken Language Understanding*, SLU) sont des alternatives efficaces aux méthodes à base de règles (Levin *et al.,* 1995, Pla *et al.,* 2001, He *et al.,* 2005, Raymond *et al.,* 2006, Lefèvre, 2007). Elles réduisent les besoins en expertise humaine et les coûts de développement tout en ayant la capacité de produire des réseaux d'hypothèses ou des listes de *n*-meilleures hypothèses. Dans un système de dialogue oral, le module de SLU est l'interface entre la reconnaissance automatique de la parole (*Automatic Speech Recognition*, ASR) et le gestionnaire de dialogue. Son rôle est d'analyser la requête du locuteur et d'en déduire une représentation de son contenu sémantique. Cette représentation permet au gestionnaire de dialogue de choisir la meilleure action à réaliser considérant le contexte courant du dialogue.

Des systèmes de SLU dans lesquels la totalité du processus de compréhension est stochastique ont été présentés dans de précédents travaux (Bonneau-Maynard *et al.,* 2005a, Raymond *et al.,* 2006, Lefèvre, 2007). Généralement conçus pour gagner en robustesse, ces systèmes raffinent progressivement les hypothèses de concepts produites. L'objectif de notre travail est de produire une information sémantique plus riche, d'une façon pertinente et généralisable. Pour ce faire, une étape de composition sémantique complémentaire est considérée afin de capturer les notions sémantiques abstraites sous-tendues par la représentation conceptuelle de base.

En l'absence de consensus sur la définition de structures de représentation sémantique adaptées à un système de dialogue oral, nous avons choisi d'utiliser le formalisme des CS et de les définir en accord avec le paradigme du projet FrameNet de Berkeley (Fillmore *et al.,* 2003). Un CS décrit une situation abstraite ou concrète impliquant des notions prédéfinies. Ceux du projet FrameNet étant pour la plupart très généralistes, des CS adaptés à la tâche visée (Meurs *et al.,* 2008) ont été définis. Un processus semi-manuel à base de règles a été développé pour fournir une annotation en CS de dialogues manuellement transcrits et annotés en concepts. Cette annotation en CS, bien qu'imparfaite, est assez fiable. Cependant, les erreurs commises sur les mots et les concepts par les systèmes automatiques de transcription et d'annotation conceptuelle doivent être prises en compte. Il est donc nécessaire de concevoir un système capable de produire des listes de *n*-meilleures hypothèses utilisables lors des étapes de validation ultérieures.

Dans cette optique, nous proposons dans cet article un système de SLU basé sur deux étapes de décodage utilisant des réseaux bayésiens dynamiques. La première étape déduit les concepts de base de la transcription des propos du locuteur (Lefèvre, 2007). Dans la seconde étape, un modèle à base de DBN réalise des inférences séquentielles sur des CS, en tenant compte de tous les niveaux d'annotation disponibles (mots et concepts). Notre postulat est qu'à partir d'un corpus d'entraînement annoté en CS, un décodage sequentiel permet de découvrir et de composer les CS associés aux propos d'un locuteur.

Le corpus MEDIA, support des expériences, est présenté dans la partie suivante de cet article. La partie 3 rappelle ensuite les notions fondamentales qui sous-tendent les CS et décrit le processus à base de règles utilisé pour fournir l'annotation de référence

en CS du corpus MEDIA. La partie 4 expose le modèle à base de DBN employé pour la composition des CS. Enfin, les résultats expérimentaux sont détaillés dans la partie 5.

2. Description des données MEDIA

Le corpus MEDIA (Bonneau-Maynard *et al.,* 2005b) est un corpus de dialogues en français issu de la simulation d'un serveur téléphonique d'informations touristiques et de réservation d'hôtels. Il a été enregistré selon le protocole du *Magicien d'Oz* (système de dialogue simulé par un opérateur humain) et contient 1250 dialogues produits par 250 locuteurs, pour une durée totale d'environ 70 heures d'enregistrement audio. Le corpus est transcrit manuellement et enrichi par une annotation également manuelle utilisant 83 concepts de base rassemblés dans un dictionnaire sémantique de concepts. Ce dictionnaire associe à un mot ou un groupe de mots une paire *concept-valeur* puis un spécifieur définissant des relations entre concepts et enfin un *mode* (affirmatif, négatif, interrogatif ou éventuel) attaché au concept. Un exemple de message annoté du corpus MEDIA est donné dans le tableau (1). La première colonne contient les groupes de mots W^c supports de chaque concept, présenté dans la seconde colonne. La troisième colonne indique le mode et la quatrième colonne fournit les spécifieurs associés aux concepts. La dernière colonne présente les valeurs normalisées des concepts c associés aux groupes de mots W^c.

W^c	*concept c*	*mode*	*specifieur*	*valeur*
Je voudrais reserver	`commande`	+ (affirmatif)		reservation
une chambre	`chambre-quantite`	+	reservation	1
pour deux nuits	`sejour-nbNuit`	+	reservation	2
à Marseille	`localisation-ville`	+	hotel	Marseille

Tableau 1. *Exemple d'annotation sémantique* MEDIA

La combinaison des spécifieurs et des concepts permet de recomposer une représentation hiérarchique de la requête du locuteur à partir de l'annotation *à plat*. Cette annotation fournit des étiquettes comparables aux constituants proposés par un analyseur sémantique de surface. Cependant, pour obtenir une représentation complète de la composition sémantique d'une proposition, l'utilisation de structures plus riches et plus complexes que les spécifieurs est nécessaire.

3. Annotation en *Cadres Sémantiques*

Les connaissances sémantiques théoriques permettent de définir des structures sémantiques adaptées à divers domaines ainsi que l'ont montré (Woods, 1975) avec les réseaux sémantiques représentant des entités/relations ou (Jackendoff, 1990) avec les structures de type fonction/argument. Une façon efficace de modéliser les connaissances sémantiques est de les représenter comme un ensemble de formules logiques sur lequel repose le processus de compréhension. Dans ce contexte, Fillmore définit les CS comme des structures cognitives associées au processus de compréhension (Fillmore, 1982, Fillmore, 1985). Chaque CS comporte des élements (CSE), rôles sémantiques, qui lui sont propres. Un CS est donc un modèle représentant des entités sémantiques et leurs propriétés (Petruck, 1996).

Le choix d'une annotation en CS dans ce travail est motivé par leur capacité à

représenter des dialogues de négociation et à s'adapter aux actions complexes du gestionnaire de dialogue. Un CS décrit une situation concrète ou abstraite impliquant ses CSE. Les mots ou groupes de mots déclenchant l'instanciation d'un CS sont ses *unités lexicales* (LU). Ces LU associent un mot (ou groupe de mots) à une signification. Le projet FrameNet de Berkeley fournit une base de données de CS pour la langue anglaise mais il n'existe pas encore de base de données comparable pour la langue française. Nous avons donc défini manuellement un ensemble de CS pour décrire nos connaissances en terme de composition sémantique adaptée au domaine du corpus MEDIA. Cette base de connaissances contient 21 CS et 86 CSE définis par des modèles composés de LU et de concepts de base (*unités conceptuelles* CU). Ces CU sont issus du dictionnaire sémantique de concepts MEDIA dès lors qu'ils peuvent être associés à un CS. Dans le cas contraire, des CU adaptés ont été définis. Le tableau 2 présente un extrait de la définition du CS `LOCATION` avec l'un de ses CSE, `location_town`.

```
<frame frname="LOCATION">
    <concept value="locate" />
    <lexical_units value="place,area" />
    <framelement fename="location_town">
      <concept value="town" />
      <specific_lexical_units value="paris,marseille..." />
    </framelement> ...
</frame>
```

Tableau 2. *Extrait de la définition du CS* MEDIA *LOCATION.*

Pour obtenir des annotations en CS sur les données d'entraînement, un processus d'annotation en deux étapes à base de règles a été développé. La première étape utilise les modèles définissant les CS pour déclencher l'instanciation de CS et de leurs CSE selon que LU et CU sont rencontrés dans les données à annoter. La seconde étape compose les CS découverts durant l'étape 1 grâce à l'application d'une série de règles logiques. CS et CSE déterminent les valeurs de vérités des prédicats. Selon ces valeurs de vérité, des CS et CSE peuvent être créés, supprimés, modifiés ou reliés. Ce dernier cas est justifié par la hiérarchie présente au sein des CS, traduite par l'aptitude de certains CSE à prendre des CS pour valeurs. Le langage de programmation logique Prolog (Colmerauer *et al.,* 1993) a été utilisé pour réaliser les inférences logiques. Un programme Prolog se compose d'une base de faits et de règles logiques décrivant les relations entre des faits potentiels. L'implémentation des règles est réalisée sous SWI-Prolog (Wielemaker, 2003). L'exemple de la règle Prolog [`do_link(RESL,L) :- is_fe(reservation_theme,RESL), is_concept_of(lodging,RESL), is_fr(lodging,L).`] illustre la création de relations : un CSE, `Reservation_Theme`, appartenant par nature au CS `RESERVE`, prend pour valeur un autre CS, `LODGING`. Les 70 règles appliquées actuellement ne prennent en compte ni les mots, ni l'ordre d'instanciation des CS lors de l'étape 1. La plupart de ces règles relient CS et CSE, instancient ceux non découverts par l'étape 1 et suppriment les redondances. Ce processus permet d'obtenir une annotation en

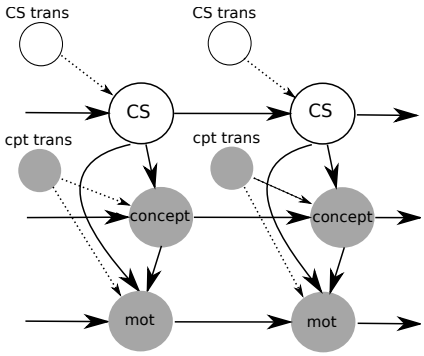

Figure 1. *Le modèle DBN pour la composition des CS.*

CS de référence pour le corpus d'entraînement sur lequel les modèles stochastiques pourront être appris.

4. Composition des *Cadres Sémantiques*

Les réseaux bayésiens dynamiques sont des modèles d'une grande flexibilité permettant de représenter des systèmes stochastiques complexes. Les DBN, utilisés dans de nombreuses tâches de modélisation de données séquentielles, obtiennent des résultats au niveau de l'état de l'art.

La figure 1 montre le modèle génératif à base de DBN utilisé pour la composition des CS dans notre système de SLU. Les sommets symbolisent les variables et les arcs signifient les dépendances conditionnelles. Sur la figure, seulement deux séquences temporelles (ou deux mots) sont indiquées. En pratique, le schéma est répété tout au long de la séquence de mots. Les sommets sont grisés lorsque les variables sont observées tandis qu'ils restent clairs pour les variables cachées. Les arcs pleins illustrent les dépendances conditionnelles entre les variables, les arcs pointillées indiquent les relations aux variables transitionnelles (modifiant les relations conditionnelles entre les autres variables). Un exemple de variable transitionnelle est donné par le sommet clair `CS trans` qui influence le sommet CS : si `CS trans` est nul, le CS est identique à son prédécesseur ; si `CS trans` est égal à un, la nouvelle valeur du CS est déterminée en fonction de la probabilité $P(f|f_{-1})$ du CS f connaissant le CS précédent f_{-1}. Toutes les variables sont observées pendant l'entraînement du modèle. Ainsi, les tables de probabilités conditionnelles associées aux arcs sont directement obtenues à partir des observations. Le calcul des valeurs de ces tables est réalisé grâce à des modèles de langage factorisés (FLM) utilisant des techniques de repli parallèle généralisé (generalized parallel backoff : GPB) (Kirchhoff *et al.,* 2007). Les FLM sont une extension des modèles de langage classiques (LM) dans laquelle les prédictions sont basées sur un ensemble de caractéristiques et non plus seulement sur les précédentes occurences de la variable. Le GPB permet d'étendre les procédures de repli standard au cas où des caractéristiques de différents types sont considérées, sans contrainte temporelle imposée (contrairement aux modèles classiques).

Dans notre modèle de composition à base de DBN, plusieurs implémentations de

FLM sont utilisées, correspondant aux arcs du graphe de la figure 1 :

– $P(F) \simeq \prod P(f|f_h)$: séquences de CS F ;

– $P(C|F) \simeq \prod P(c|c_h, f)$, repli GPB dans l'ordre $\{c_h, f\}$: séquences de concepts C conditionnées par les CS F ;

– $P(W|C, F) \simeq \prod P(w|w_h, c, f)$, repli GPB dans l'ordre $\{w_h, c, f\}$: séquences de mots W conditionnées par les concepts C et les CS F.

L'indice h représente un historique qui peut varier en fonction de la longueur des séquences prises en compte par les modèles ($\{-1\}$ pour les 2-grammes, $\{-1, -2\}$ pour les 3-grammes, etc.). Toutes les expériences exposées ont été réalisées avec GMTK (Bilmes *et al.,* 2002), outil logiciel de calcul et de manipulation des modèles graphiques, et SRILM (Stolcke, 2002), outil logiciel pour les modèles de langage.

Actuellement, notre modèle DBN ne décode que les CS (i.e. les CSE ne sont pas considérés). Pour prendre en compte les situations de chevauchement (cas où plusieurs CS peuvent être associés aux mêmes mots ou concepts), des classes de CS "composés" sont considérées lors du processus de décodage puis les CS sont séparés par la suite. Les séquences de `mots`, `concepts` et `cpt trans` associées sont des variables observées pour le décodage des CS : elles ont été décodées par les modules d'ASR et de SLU. Etant donnée la faible densité de données, les probabilités conditionnelles utilisées dans le modèle sont limitées à celles obtenues par un FLM 2-grammes.

5. Expériences et Résultats

Pour évaluer les performances du système de composition des CS utilisant les DBN, on définit un ensemble de données de test. Pour des raisons de temps et de coûts, seulement 15 dialogues (225 tours de parole du locuteur) ont été annotés en CS par un expert humain. Le système d'annotation en deux étapes à base de règles a été utilisé pour produire une annotation en CS sur le corpus MEDIA (transcriptions et annotation conceptuelle manuelles), données de test exclues. Les FLM utilisés dans le modèle DBN ont été entraînés sur ces données annotées.

Les expériences sont menées sur l'ensemble de test dans trois conditions différentes, fonctions du type des données initiales : (i) la référence (REF) où les propos du locuteur sont transcrits et annotés en concepts et CS manuellement ; (ii) le cas SLU où les concepts sont décodés à partir des transcriptions manuelles des propos du locuteur, grâce à un modèle de SLU à base de DBN (Lefèvre, 2006) ; (iii) le cas ASR+SLU où les séquences de mots proviennent de la meilleure hypothèse générée par le système d'ASR (Barrault *et al.,* 2008) et le décodage conceptuel utilise ces hypothèses. Dans le tableau 3, les taux d'erreurs en mots et en concepts sont précisés pour les trois types de données initiales. L'évaluation du système à base de règles fournit une base de comparaison. Le tableau 3 rassemble les résultats des deux systèmes en termes de précision, rappel et F-mesure. La précision est le nombre de CS corrects proposés par le système divisé par le nombre total de CS proposés par le système. Le rappel est le nombre de CS corrects proposés par le système divisé par le nombre total de CS contenus dans l'annotation de référence. Pour les deux systèmes, seuls les CS sont pris en compte : ni leurs éléments, ni l'ordre de leur instanciation ne sont considérés.

Systèmes	Données	REF	SLU	ASR + SLU
	WER	0.0	0.0	14.8
	CER	0.0	10.6	24.3
	P	0.94	0.92	0.88
à base de règles	R	0.92	0.87	0.80
	F-m	0.93	0.89	0.84
	P	0.91	0.91	0.82
à base de DBN	R	0.89	0.79	0.76
	F-m	0.90	0.85	0.79

Tableau 3. *Précision (P), Rappel (R) et F-mesure (F-m) obtenus sur les données de test du corpus* MEDIA *pour les systèmes de composition des CS. Taux d'erreurs en mots (WER) et en concepts (CER) (%) sur les ensembles de test pour les trois types de données : référence (REF), à partir des transcriptions manuelles (SLU) et à partir de la meilleure hypothèse fournie par le module d'ASR (ASR+SLU).*

La F-mesure est la moyenne harmonique standard de la précision et du rappel.

Les résultats du tableau 3 montrent que le système utilisant les DBN obtient des résultats comparables à celui basé sur des règles. La faible différence entre les deux systèmes (environ 0,05 point d'écart entre les F-mesures) reste constante dans les trois conditions d'expérimentation. L'ensemble de test étant encore de taille modeste, l'amplitude de l'intervalle de confiance est d'environ 0,03. Les différences observées sont donc statistiquement peu significatives. Une F-mesure de 0,93 pour le système à base de règles, dans le cas le moins bruité, confirme la fiabilité du processus à base de règles. De plus, il est intéressant de souligner que le système utilisant les DBN fournit en une seule étape ce qu'un expert humain conçoit en deux étapes : obtention d'un premier ensemble de CS par comparaison des données avec les définitions des CS puis composition des CS de cet ensemble à l'aide de règles logiques pour prendre en compte leurs relations dans les propos du locuteur.

6. Conclusion

Cet article présente un processus d'interprétation stochastique pour la composition de CS utilisant les DBN. Ce processus, dédié à l'interprétation du langage oral, permet de déduire des CS à partir des séquences de mots et de concepts de base présentes dans les propos du locuteur. Les résultats expérimentaux, obtenus sur le corpus de dialogue MEDIA, montrent que les performances du modèle à base de DBN sont comparables à celles d'un système expert à base de règles.

L'approche proposée est efficace pour découvrir automatiquement les CS sous-tendus par les phrases du locuteur. L'enrichissement du modèle à base de DBN par la prise en compte des CSE sera la prochaine étape de ce travail. Grâce à la grande flexibilité des DBN en terme de représentation des probabilités, cela consistera essentiellement à ajouter une nouvelle variable dans le graphe, en lui associant les probabilités conditionnelles appropriées.

De plus, les n-meilleures hypothèses de séquences de mots et de concepts fournies par les modules d'ASR et de SLU pourront être utilisées pour obtenir les n-meilleures hypothèses de CS et leurs indices de confiance.

7. Bibliographie

Barrault L., Servan C., Matrouf D., Linarès G., Mori R. D., « Frame-Based Acoustic Feature Integration for Speech Understanding », *IEEE ICASSP*, 2008.

Bilmes J., Zweig G., « The graphical models toolkit : An open source software system for speech and time-series processing », *IEEE ICASSP*, 2002.

Bonneau-Maynard H., Lefèvre F., « A 2+1-level stochastic understanding model », *IEEE ASRU*, 2005a.

Bonneau-Maynard H., Rosset S., Ayache C., Kuhn A., Mostefa D., the Media consortium, « Semantic annotation of the MEDIA corpus for spoken dialog », *ISCA Eurospeech*, 2005b.

Colmerauer A., Roussel P., « The birth of Prolog », *HOPL-II : The second ACM SIGPLAN conference on History of programming languages*, p. 37-52, 1993.

Fillmore C., *Frame Semantics*, Linguistics in the Morning Calm, Seoul, 1982.

Fillmore C., « Frames and the Semantics of Understanding », *Quaderni di Semantica*, vol. VI, n° 2, p. 222-254, 1985.

Fillmore C., Johnson C., Petruck M., « Background to FrameNet », *International Journal of Lexicography*, vol. 16.3, p. 235-250, 2003.

He Y., Young S., « Spoken Language Understanding using the Hidden Vector State Model », *Speech Communication*, vol. 48(3-4), p. 262-275, 2005.

Jackendoff R., « Semantic Structures », *The MIT Press, Cambridge Mass.*, 1990.

Kirchhoff K., Bilmes J., Duh K., Factored Language Models Tutorial, Technical Report n° UWEETR-2007-0003, Dept. of EE, U. Washington, June, 2007.

Lefèvre F., « A DBN-based multi-level stochastic spoken language understanding system », *IEEE Workshop on SLT*, 2006.

Lefèvre F., « Dynamic Bayesian Networks and Discriminative Classifiers for Multi-Stage Semantic Interpretation », *IEEE ICASSP*, 2007.

Levin E., Pieraccini R., « Concept-based Spontaneous Speech Understanding System », *ESCA Eurospeech*, Madrid, p. 555-558, 1995.

Meurs M.-J., Duvert F., Béchet F., Lefèvre F., Mori R. D., « Semantic Frame Annotation on the French MEDIA corpus », *LREC*, 2008.

Petruck M., « Frame semantics », *Handbook of Pragmatics*, 1996.

Pla F., Molina A., Sanchis E., Segarra E., Garcia F., « Language Understanding using Two-level Stochastic Models with POS and Semantic Units », *LNCS series*, vol. 2166, p. 403-409, 2001.

Raymond C., Béchet F., Mori R. D., Damnati G., « On the use of finite state transducers for semantic interpretation », *Speech Communication*, vol. 48 :3-4, p. 288-304, 2006.

Stolcke A., « SRILM an extensible language modeling toolkit », *IEEE ICASSP*, 2002.

Wielemaker J., « An overview of the SWI-Prolog Programming Environment », *in* , F. Mesnard, , A. Serebenik (eds), *Proceedings of the 13th International Workshop on Logic Programming Environments*, Katholieke Universiteit Leuven, Heverlee, Belgium, p. 1-16, december, 2003. CW 371.

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