Un Analyseur Sémantique pour le DHM
Démonstration

Jérôme Lehuen (1), Thierry Lemeunier (2)

LIUM - Université du Maine
Avenue Laënnec, 72085 Le Mans Cedex 9
(1) Jerome.Lehuen@lium.univ-lemans.fr
(2) Thierry.Lemeunier@lium.univ-lemans.fr

1 Introduction

Le système présenté est un analyseur sémantique destiné à être intégré dans un système de dialogue homme-machine avec entrée vocale. Afin de ne pas faire reposer l'interprétation sémantique sur le succès d'une analyse syntaxique, nous avons cherché à intégrer des aspects syntaxiques et sémantiques au sein d'un même modèle, ce qui permet plus facilement de faire collaborer ces deux dimensions. De plus, l'analyseur permet les analyses partielles, ainsi que la génération d'hypothèses lexicales et structurelles. Il repose sur une grammaire syntaxico-sémantique contrôlée par un processus de type analyse tabulaire (Kay, 1986) qui permet de contourner certaines structures agrammaticales, ou non décrites dans la grammaire.

2 Modélisation

Les connaissances se présentent sous la forme d'un lexique dans lequel les entrées sont définies à la fois syntaxiquement et sémantiquement. Une lexie est caractérisée par un nom (forme normalisée), par un système d'offres et d'attentes sémantiques, et par au moins une forme d'usage langagière (Figure 1). Les offres et les attentes sont des catégories ou des traits sémantiques, déterminés à partir d'une ontologie issue de la tâche applicative. Les formes d'usage sont des structures syntaxiques composées de mots (symboles terminaux) et/ou de relations de dépendance faisant référence aux attentes (symboles non-terminaux), comme par exemple « A1 chambre A2 » où A1 et A2 font référence à des attentes de la lexie [chambre]. On obtient ainsi une sorte de grammaire non contextuelle utilisable pour l'analyse, et éventuellement pour la génération (Lehuen, 2008).

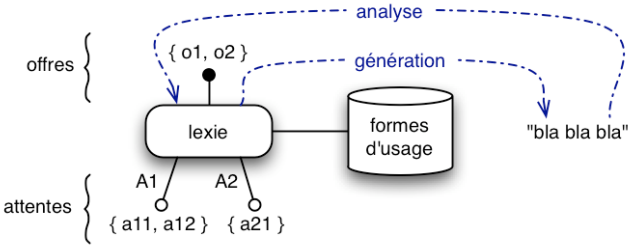


Figure 1 : Modèle de lexie à deux attentes

La représentation sémantique d'un énoncé consiste alors en une structure de dépendances (Tesnière, 1959) qui couvre les mots ou expressions correspondant à des formes d'usage reconnues. Les nœuds du graphe – les *granules* – correspondent à des lexies instanciées dans le contexte de l'énoncé (Figure 2).

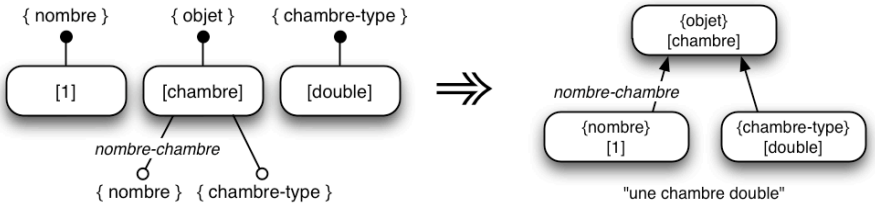


Figure 2 : Des lexies à la structure de granules

La lexie [chambre] comporte une attente associée à un rôle. Le rôle permet de contextualiser un granule fils en fonction de son père. Ainsi, un granule possédant le trait {nombre} dans le contexte d'un granule [chambre] est un *nombre-chambre*.

3 Implémentation et résultats

L'algorithme d'analyse tabulaire repose sur l'algorithme RETE (Forgy, 1982) qui intervient dans l'implémentation de systèmes à base de faits et de règles de production, tel que CLIPS[1]. Pratiquement, le lexique codé en XML est d'abord traduit sous la forme de règles CLIPS à l'aide de transformations XSLT, puis compilé par l'algorithme RETE. Cet algorithme, qui tend à sacrifier la mémoire au profit de la vitesse, s'est révélé particulièrement efficace pour implémenter notre algorithme. Les 3000 énoncés de 8 mots en moyenne du corpus MEDIA (Bonneau-Maynard, 2005) ont été analysés en moins de 4mn, soit 80ms par énoncé sur un Dual Core 2Ghz (ce temps évoluant linéairement par rapport au nombre de mots). Nos résultats sur le corpus MEDIA sont plutôt satisfaisants avec un CER (*Concept Error Rate*) égal à 34,7 par rapport à l'annotation manuelle en « full 4 modes ». Le tableau suivant place notre taux d'erreurs par rapport aux participants à la campagne MEDIA-EVALDA de 2005 :

LIMSI-1	LIMSI-2	LIUM	LORIA	VALORIA	LIA
29,0	30,3	34,7	36,3	37,8	41,3

4 Démonstration

Notre démonstration consiste à illustrer la grammaire et le fonctionnement de l'analyseur à l'aide d'un lexique et d'un corpus simplifiés. Les points abordés sont les suivants :

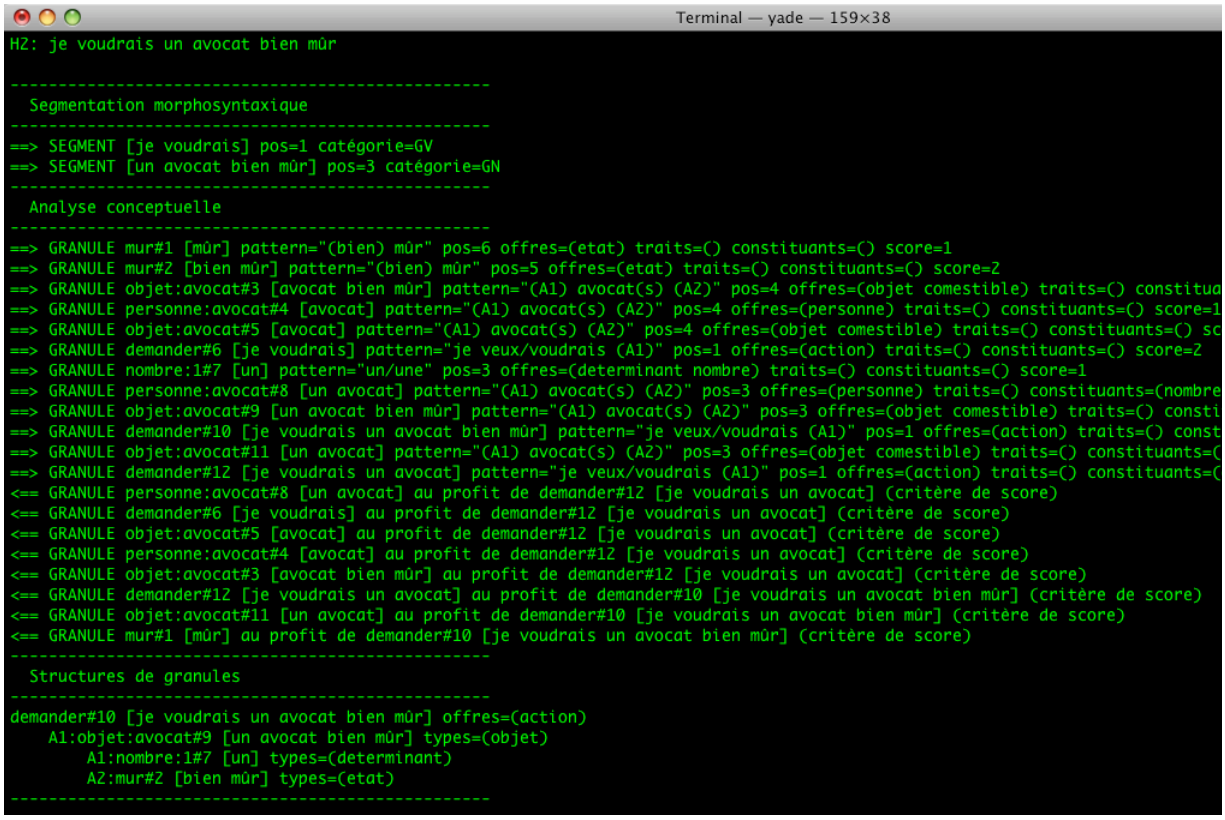
1. Le modèle des connaissances (granules, structures de granules, rôles) ;
2. La résolution des ambiguïtés lexicales (utilisation des traits sémantiques) ;
3. La génération des hypothèses lexicales (utilisation des formes d'usage) ;
4. La génération des hypothèses structurelles (utilisation des traits sémantiques) ;
5. La représentation des questions et des négations.

[1] CLIPS (*C Language Integrated Production System*) est un outil de construction de systèmes à base de règles.

Chaque point fait l'objet d'un mini scénario qui illustre la formalisation des connaissances, les principes d'analyse mis en œuvre, ainsi que la représentation obtenue. Par exemple, le point n°2 (résolution des ambiguïtés lexicales) est illustré à l'aide des phrases suivantes :

- Un avocat — ambiguïté personne/fruit non résolue
- Un avocat bien mûr — ambiguïté résolue par le fils (adjectif)
- Un avocat très compétent — ambiguïté résolue par le fils (adjectif)
- Je mange un avocat — ambiguïté résolue par le père (verbe)
- Je mange un avocat très compétent — ambiguïté personne/fruit non résolue

La trace des inférences permet de suivre les différentes étapes de l'analyse sémantique :



Références

BONNEAU-MAYNARD H., ROSSET S., AYACHE C., KUHN A., MOSTEFA D. (2005). Semantic Annotation of the French MEDIA dialog corpus. *Proceedings of Eurospeech 05*, Lisboa.

FORGY, C.L. (1982). Rete: A Fast Algorithm for the Many Pattern/Many Object Pattern Match Problem. *Artificial Intelligence* (19), 17-37.

KAY M. (1986). Algorithms chemata and data structures in syntactic processing. *Readings in natural language processing*. San Francisco, CA, USA : Morgan Kaufmann Publishers Inc.

LEHUEN J. (2008). Un modèle de langage pour le DHM : la Grammaire Sémantique Réversible (poster), *Actes de TALN'08*, 9-13 juin 2008, Avignon (France).

TESNIÈRE L. (1959). *Éléments de syntaxe structurale*, Paris, Klincksiek.