Construction automatique d'une interface syntaxe / sémantique utilisant des ressources de large couverture en langue anglaise

François-Régis Chaumartin

Lattice/Talana – Université Paris 7
fchaumartin@linguist.jussieu.fr

Résumé

Nous décrivons ici une approche pour passer d'une représentation syntaxique (issue d'une analyse grammaticale) à une représentation sémantique (sous forme de prédicats). Nous montrons ensuite que la construction de cette interface est automatisable. Nous nous appuyons sur l'interopérabilité de plusieurs ressources couvrant des aspects d'ordre syntaxique (Link Grammar Parser), lexical (WordNet) et syntaxico-sémantique (VerbNet) de la langue anglaise. L'utilisation conjointe de ces ressources de large couverture permet d'obtenir une désambiguïsation syntaxique et lexicale au moins partielle.

Mots-clés : compréhension de texte, analyse syntaxique, désambiguïsation, interface syntaxe-sémantique, reconnaissance de schémas, logique des prédicats, unification, rôles thématiques, contraintes de sélections, VerbNet, WordNet, Link Grammar Parser.

Abstract

We describe a way to transform a syntactic structure (generated by a syntactic parser for English) into a semantic form (in the form of predicates). We then show that the construction of such an interface can be automated. Our approach is based on the interoperability between several resources, covering syntactical (Link Grammar Parser), lexical (WordNet) and semantic (VerbNet) aspects of English. The joint use of these broad-coverage resources leads to a lexical and syntactical disambiguation (at least partially).

Keywords: text understanding, parsing, disambiguation, syntax/semantic interface, pattern recognition, predicate calculus, unification, thematic roles, selectional restrictions, VerbNet, WordNet, Link Grammar Parser.

1. Introduction

Notre projet de thèse vise à extraire des connaissances d'une encyclopédie en langue anglaise. Dans ce contexte, nous souhaitons disposer d'une représentation sémantique d'un texte. Notre démarche consiste, partant d'une forme syntaxique du texte (l'arbre de dépendance obtenu en sortie d'un analyseur syntaxique), à passer par application d'heuristiques successives vers une forme sémantique ; sa représentation est un graphe sémantique dont les nœuds sont des acceptions désambiguïsées d'unités lexicales. Pour le cadre théorique, nous nous inspirerons de la Théorie Sens-Texte et plus particulièrement l'interface syntaxe/sémantique de la Grammaire d'Unification Sens-Texte (Kahane, 2002). Nous présentons ici l'une des heuristiques permettant de désambiguïser un verbe.

Parmi les travaux qui utilisent des ressources comparables aux nôtres, citons Shi et Mihalcea (2005) qui revendique la construction d'un analyseur sémantique robuste en langue anglaise, en combinant les ressources WordNet, VerbNet et FrameNet (Baker, 1998). Dzikovska (2004) présente un analyseur sémantique multi domaines, et définit une algèbre sur les contraintes de sélection, ainsi qu'une méthode d'apprentissage des contraintes de sélection à partir d'un corpus. De nombreux autres articles et projets couvrent ce domaine.

L'originalité de notre approche réside d'une part en l'intégration de ressources de large couverture sur la langue anglaise, d'autre part en une démarche de génie logiciel visant à industrialiser la construction de l'analyseur sémantique. Nous commencerons par effectuer des rappels sur trois notions fondamentales (classes de verbes, rôles thématiques et contraintes de sélection). Nous décrirons ensuite chaque ressource utilisée et sa mise en œuvre. Enfin, nous détaillerons les mécanismes utilisés.

2. Rappels sur trois notions fondamentales

Une **classe de verbes** est un regroupement de verbes qui partagent un même comportement syntaxique et sémantique et qui, de ce fait, connaissent les mêmes constructions typiques. Par exemple, la classe de verbes "**murder**" regroupe plusieurs verbes tels que : *to assassinate* ('assassiner'), *to eliminate* ('éliminer'), *to execute* ('exécuter'), etc. Les verbes membres d'une classe ne sont pas forcément tous synonymes entre eux : la classe "**give**" regroupe 'donner', 'louer', 'prêter', 'rendre', 'restituer'…, lesquels partagent les mêmes constructions.

Les **rôles thématiques** (Gruber, 1965 ; Fillmore, 1968 ; Jackendoff, 1972) font référence à la relation sémantique sous-jacente entre un prédicat et ses arguments. Chaque argument du verbe remplit un rôle thématique. Il peut être, par exemple, *Agent, Patient, Thème, Instrument, Source*… de l'action ou de l'événement décrit par le verbe. Ces rôles sont indépendants de la construction syntaxique ; par exemple, dans « Jean frappe Marie » ou « Marie est frappée par Jean », « Marie » est *Patient* et « Jean » est *Agent* de l'action.

Les rôles thématiques peuvent avoir des **contraintes de sélection**, qui en restreignent les réalisations possibles (par exemple, l'*Agent* de "**murder**" doit être *Animé)*. Les contraintes de sélection sont organisées selon un graphe d'héritage ; par exemple, les ancêtres successifs d'*Humain* sont (dans VerbNet) *Animé*, *Naturel* et *Concret*. Notre analyseur sémantique utilise cette caractéristique pour désambiguïser le sens des mots, en établissant une correspondance entre graphe de mots du lexique et hiérarchie des contraintes de sélection.

3. Description des ressources utilisées

3.1. Analyse syntaxique

Le *Link Grammar Parser* (Sleator et Temperley, 1991) est un analyseur syntaxique de la langue anglaise, basé sur la syntaxe de dépendance. Cet analyseur offre de bonnes performances et semble robuste. Il traite avec succès des phrases complexes, en tolérant la présence de mots inconnus. Partant d'une phrase fournie en entrée, cet analyseur produit un ou plusieurs graphes orientés acycliques de dépendances, qui consistent en un ensemble de liens typés reliant des paires de mots. Les **nœuds** du graphe sont les mots de la phrase ; certains d'entre eux ont un suffixe qui indique la partie du discours (nom, verbe, adjectif, adverbe, préposition, etc.). Des **arcs étiquetés** relient les nœuds du graphe, chaque étiquette précisant un rôle grammatical (**D** pour déterminant-nom, **S** pour sujet-verbe, etc.).

3.2. Analyse lexicale

Deux ressources lexicales ont été fusionnées pour être facilement utilisées ensemble :

1. *WordNet* (Miller, 1995) est un projet mené depuis 1985 à Princeton. WordNet offre un réseau sémantique de la langue anglaise. Les nœuds sont constitués par des ensembles de synonymes (ou *synsets*), correspondant au sens d'un ou plusieurs lemmes. Un *synset* est défini d'une façon différentielle par les relations qu'il entretient avec les sens

voisins. WordNet nous sert à déterminer les différents sens d'un mot donné et à chercher lesquels vérifient certaines contraintes de sélection.

2. *eXtended WordNet* (XWN) est un projet de l'Université de Dallas, qui enrichit les relations de WordNet. XWN ajoute une forme logique à chaque définition textuelle de *synset*. XWN nous sert à déterminer si un nom possède un attribut particulier.

3.3. Analyse sémantique

VerbNet est un lexique des classes de verbes anglais. C'est un projet mené sous l'impulsion de Martha Palmer, d'abord à l'Université de Pennsylvanie, puis à l'Université de Boulder. *VerbNet* regroupe par classe les verbes partageant les mêmes comportements syntaxiques et sémantiques. C'est un prolongement des travaux de Levin (1993). VerbNet 1.5 distingue 192 classes de verbes qui regroupent 3880 sens de verbes. On peut en trouver une description dans Kipper-Schuler (2003). Cette ressource décrit plusieurs constructions typiques (des « *frames* ») des verbes membres de la classe. La sémantique de l'action est également précisée. Des sous-classes permettent de décrire d'éventuelles spécialisations d'une classe. La ressource pour les verbes français la plus proche nous semble être le lexique-grammaire du LADL dont le principe est décrit dans Gross (1975).

Nous utilisons VerbNet en deux temps, pour la phase d'analyse sémantique. Dans une étape préparatoire, l'analyseur sémantique traduit les descriptions de classes de verbes (stockées en XML) en programmes en langage PROLOG. Lors de l'analyse effective d'un texte, ces programmes utilisent le mécanisme d'unification pour identifier des schémas, de façon à reconnaître les constructions typiques de verbes dans une phrase fournie en entrée.

4. Correspondance entre rôles thématiques et sommet de la hiérarchie des noms

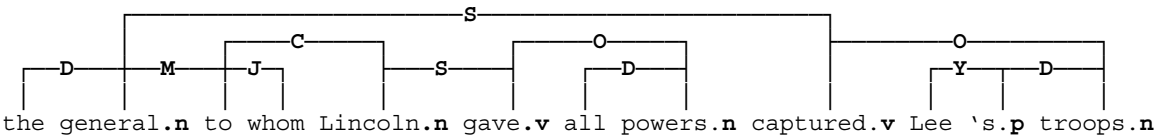
L'un des points essentiels dans la réalisation de notre interface syntaxe/sémantique est l'établissement d'une correspondance entre les rôles thématiques de VerbNet et le haut de la hiérarchie des noms dans WordNet. Nous avons défini une centaine de règles de correspondance, associant à chaque contrainte de sélection des *synsets* de WordNet. Précisons ces règles à travers un exemple. Dans la classe de verbes "**poke**" (« enfoncer »), le rôle *Instrument* a une contrainte de sélection *Pointu*. Nous vérifions deux types de règles :

1. Si un concept hérite d'un concept ancêtre particulier défini plus haut dans la hiérarchie des noms de WordNet : nous considérons par exemple comme *Pointu* tout héritier des noms 'aiguille' ("*needle*") ou 'clou' ("*nail*").

2. Si la définition textuelle d'un concept est qualifiée par un attribut particulier : nous vérifions dans ce cas que la définition contient l'adjectif 'pointu' ("*pointed*") ou 'aiguisé' ("*sharp*"). De cette façon, "*claw*" ('griffe' d'animal), qui a pour définition "a ***sharp*** *curved horny process on the toe of a bird or some mammals or reptiles*", sera reconnu comme compatible avec la contrainte de sélection *Pointu*.

5. Exemple de mise en œuvre de l'interface syntaxe / sémantique

Partons d'une phrase extraite d'un article sur la guerre de Sécession : "*the general to whom Lincoln gave all powers captured Lee's troops*" (« le général à qui Lincoln avait donné tous les pouvoirs captura les troupes de Lee »). Notre donnée de départ est le graphe de dépendances fourni par le *Link Grammar Parser* :



Nous le représentons par un ensemble de prédicats PROLOG `word` (mot) et `link` (lien) :

```
word(1, "the", article(nil)).
word(2, "general", noun("general", baseform)).
…
link(2, d, 1, 2). /* the - general */
link(3, s, 2, 9). /* general - captured */
…
```

Les informations fournies par VerbNet permettent d'identifier ici deux constructions typiques de verbe, ainsi que les contraintes de sélection portant sur les rôles thématiques :

the **general***(Recipient)* to whom **Lincoln***(Agent)* **gave***(Verb)* all **powers***(Theme)* captured Lee's troops

the **general***(Agent)* to whom Lincoln gave all powers **captured***(Verb)* Lee's **troops***(Theme)*

Remarquons que la phrase contenait une relative ; initialement, le système était limité à la reconnaissance des constructions simples où sujet, verbe et compléments se suivent d'une façon linéaire ; dans un corpus réel, les phrases n'étant pas aussi triviales, nous avons étendu le système pour identifier des constructions grammaticales plus complexes.

Les constructions reconnues par l'interface syntaxe / sémantique donnent alors les prédicats :

```
frame_give(6 /* Verb=gave */,
  5 /* Agent=Lincoln */, 8 /* Theme=powers */, 2 /* Recipient=general */).
frame_capture(9 /* Verb=captured */,
  2 /* Agent=general */, 12 /* Theme=troops */).
has_constraint(2 /* general */, animate).
has_constraint(5 /* Lincoln */, animate).
```

Dans notre lexique (WordNet), chaque mot a plusieurs sens possibles. Muni des informations de contraintes de sélection, l'analyseur restreint la liste des sens des mots acceptables dans le contexte et effectue une **désambiguïsation lexicale**, totale ou partielle selon le cas. Dans notre exemple, le verbe "give" possède 44 sens, mais seuls 6 d'entre eux sont compatibles avec le *frame*. De même, le verbe "capture" a 6 sens, dont un seul acceptable dans le contexte.

La **désambiguïsation syntaxique** (c'est-à-dire le choix de l'interprétation syntaxique qui semble optimale) est obtenue en conservant, dans la forêt produite par le *Link Grammar Parser*, le graphe syntaxique contenant le plus grand nombre de rôles thématiques reconnus.

6. Détection de constructions verbales typiques

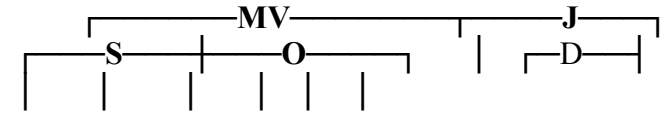
Nous utilisons un mécanisme de reconnaissance de schémas pour détecter des constructions typiques de verbes, afin de passer d'une représentation syntaxique à une représentation sémantique. Pour ce faire, nous traduisons la description de la syntaxe d'une classe de verbes de VerbNet en un programme PROLOG, capable d'extraire du graphe syntaxique créé par le *Link Grammar Parser* chaque verbe prédicat avec ses arguments (ses rôles thématiques).

6.1. Correspondance entre concepts de VerbNet et code PROLOG

VerbNet associe un exemple à chaque *frame*. La description de sa syntaxe donne le cadre de sous-catégorisation permettant d'identifier une construction typique. Notre idée maîtresse consiste à utiliser cet exemple, en établissant une correspondance entre [a] la déclaration de

syntaxe du *frame* fournie par VerbNet, et [b] la syntaxe de la phrase d'exemple du *frame* analysée par le *Link Grammar Parser*. Par exemple, le deuxième *frame* de la classe de verbe "**murder**" se compose des éléments consécutifs : *Agent*, *Verbe*, *Patient*, "with", *Instrument*.

Examinons l'analyse syntaxique de la phrase d'exemple "*Brutus killed Caesar with a knife*" :

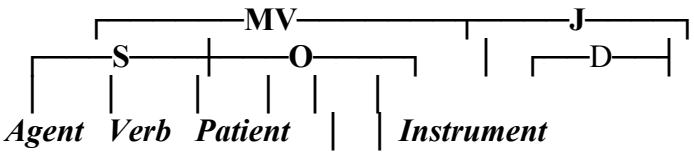


Brutus **killed.v** Caesar.n **with** a knife.n
mot.1 mot.2 mot.3 mot.4 mot.6

Après avoir identifié le verbe, nous conservons les liens partant et arrivant au verbe et aux autres éléments littéraux explicitement cités :

```
link(2, s, 1, 2). /* Brutus - killed */
link(4, o, 2, 3). /* killed - Caesar */
link(3, mv, 2, 4)./* killed - with */
link(5, j, 4, 6). /* with - knife */
```

Il ne nous reste plus qu'à aligner l'analyse de la phrase d'exemple et la description du *frame* :



Brutus killed.v Caesar.n with a knife.n

6.2. Exemple : codage du prédicat identifiant le *frame* 2 de "*murder*"

En disposant de cette table de correspondance, nous pouvons écrire le prédicat PROLOG qui reconnaît le *frame*. On identifie un schéma avec les quatre liens correctement reliés entre eux et la préposition "*with*".

```
frame_murder_2(L2:L4:L3:L5, NAgent, NVerb, NPatient, NInstrument) :-
link(L2, s, NAgent, NVerb),
link(L4, o, NVerb, NPatient),
link(L3, mv, NVerb, NWith),
word(NWith, "with", preposition),
link(L5, j, NWith, NInstrument),
…
```

Nous devons vérifier des contraintes :

```
/* Patient doit apparaître avant Instrument */
NPatient < NInstrument,

/* Agent, Instrument, Patient ont des contraintes de sélection */
isAgent_murder_42_1(NAgent),
isPatient_murder_42_1(NPatient),
isInstrument_murder_42_1(NInstrument),

/* La forme de base du verbe doit être membre de la classe de verbe */
word(NVerb, _VerbDerivedForm, verb(Verb, VerbForm)),
isMember_murder_42_1(Verb, WN).
…
```

Il nous reste à coder la prise en compte des contraintes de sélection sur les rôles thématiques (*Agent* et *Patient* sont *Animé* ; *Instrument* n'a pas de contrainte) :

```
isAgent_murder_42_1(NAgent) :- testSelRestr(animate, NAgent), !.
isPatient_murder_42_1(NPatient) :- testSelRestr(animate, NPatient), !.
isInstrument_murder_42_1(NInstrument) :- .
```

7. Automatisation de cette démarche

Nous avons vu comment traduire manuellement un fichier de VerbNet en un programme PROLOG équivalent. Nous avons industrialisé ce processus, en écrivant un compilateur qui, à partir des 192 fichiers XML de VerbNet version 1.5, génère automatiquement 31 600 lignes de code PROLOG identifiant les constructions verbales typiques.

Nous avons ainsi pu automatiser la traduction de 90 % des classes de verbes décrites dans VerbNet. L'un des points délicats restant à couvrir concerne les verbes contenant des arguments symétriques qui ont alors deux rôles tels qu'*Acteur1* et *Acteur2*.

8. Conclusion

Cet article montre que fédérer des ressources de large couverture permet de construire un analyseur sémantique robuste. Il met aussi en avant le double intérêt (en termes de **qualité du code** et de **réutilisation**) de générer automatiquement l'analyseur sémantique à partir des données des ressources.

9. Remerciements

Merci pour leurs conseils et relecture à Sylvain Kahane (Paris 10) et Benoît Habert (Paris 10).

Références

BAKER C., FILLMORE C., LOWE J. (1998). « The Berkeley FrameNet project ». In *Proceedings of 17th international conference on Computational linguistics.*

DZIKOVSKA M. (2004). *A practical semantic representation for Natural Language Parsing.* Ph.D. Thesis, University of Rochester.

FILLMORE C. (1968). *The case for case.* In Bach et Harms (éds), *Universals in Linguistic Theory.* Holt, Rinehart and Winston, New York : 1-88.

GROSS M. (1975). *Méthodes en syntaxe*. Hermann, Paris.

GRUBER J.S. (1965). *Studies in lexical relations*. Doctoral Dissertation. MIT, Cambridge.

JACKENDOFF R. (1972). *Semantic interpretation in generative grammar*. MIT Press, Cambridge.

KAHANE S. (2002). *Grammaire d'Unification Sens-Texte : Vers un modèle mathématique articulé de la langue*. HDR, Université Paris 7.

KIPPER-SCHULER K. (2003). *VerbNet : a broad coverage, comprehensive, verb lexicon.* Ph.D. Thesis, University of Pennsylvania.

LEVIN B. (1993). *English Verb Classes and Alternation : A Preliminary Investigation*. University of Chicago Press, Chicago.

MILLER G. (1995). « Wordnet : A lexical database ». In *Proceedings of ACM 38* : 39-41.

SHI L., MIHALCEA R. (2005). « Putting Pieces Together: Combining FrameNet, VerbNet and WordNet for Robust Semantic Parsing ». In *Proceedings of CICLing 2005. Mexico.*

SLEATOR D., TEMPERLEY D. (1991). « Parsing English with a Link Grammar ». In *Actes de Third International Workshop on Parsing Technologies*.

Ressources

WordNet – http://wordnet.princeton.edu

eXtended WordNet – http://xwn.hlt.utdallas.edu

VerbNet – http://verbs.colorado.edu/~kipper/verbnet.html

Link Grammar Parser – http://bobo.link.cs.cmu.edu/link