Annotation de textes par extraction d'informations lexico-syntaxiques et acquisition de schémas conceptuels de causalité

Laurent Alamarguy*, Rose Dieng-Kuntz*, Catherine Faron-Zucker**

*ACACIA, INRIA Sophia Antipolis
{Laurent.Alamarguy, Rose.Dieng}@sophia.inria.fr
**MAINLINE, I3S, Sophia Antipolis
faron@essi.fr

Résumé. Nous présentons la méthode INSYSE (Interface Syntaxe SEmantique) pour l'annotation de documents textuels. Notre objectif est de construire des annotations sémantiques de ces résumés pour interroger le corpus sur la fonction des gènes et leurs relations de causalité avec certaines maladies. Notre approche est semi-automatique, centrée sur (1) l'extraction d'informations lexico-syntaxiques à partir de certaines phrases du corpus comportant des lexèmes de causation, et (2) l'élaboration de règles basées sur des grammaires d'unification permettant d'acquérir à partir de ces informations des schémas conceptuels instanciés. Ceux-ci sont traduits en annotations RDF(S) sur la base desquelles le corpus de textes peut être interrogé avec le moteur de recherche sémantique Corese.

1 Introduction

Lors de la constitution d'une mémoire de communauté en génomique fonctionnelle, la notion de causalité est centrale pour appréhender certaines corrélations. Dans le cadre du *web sémantique* l'automatisation de cette tâche doit permettre, à partir de données hétérogènes, de détecter et générer de nouvelles représentations conceptuelles traduisant cette notion.

Nous présentons une méthode semi-automatique d'annotation de documents textuels basée sur l'acquisition de schémas conceptuels[1] à partir de l'extraction de structures lexico-syntaxiques ; elle est baptisée INSYSE - pour INterface SYntaxe SEmantique. Cette méthode est appliquée à un corpus de 5000 résumés médicaux issus de Medline et traitant de maladies du système nerveux central et des interactions des gènes dans ces maladies. Notre objectif est de construire des annotations sémantiques de ces résumés qui permettent d'interroger le corpus sur la fonction des gènes et leurs relations de causalité avec certaines maladies pour ainsi constituer une mémoire de communauté.

Nous présentons dans cet article les différentes étapes de la méthode INSYSE : la partie suivante est consacrée à l'extraction d'informations lexico-syntaxiques à partir de certaines phrases comportant des lexèmes de causation ; la partie 3 est dédiée à l'élaboration de règles basées sur des grammaires d'unification qui permettent d'extraire des informations lexico-syntaxiques des schémas conceptuels instanciés. La partie 4 décrit comment ces schémas sont traduits en annotations RDF(S) sur la base desquelles le corpus pourra être interrogé à l'aide du moteur de recherche sémantique Corese (Corby et al. 2004). Nous comparons dans

[1] Un schéma conceptuel non instancié constituant de fait un *template* d'annotation.

la partie 5 la méthode INSYSE aux autres approches existantes pour l'annotation de documents textuels. Nous concluons cet article en suggérant les perspectives de notre travail.

2 Extraction de lexèmes à partir de textes

La sélection des phrases des résumés du corpus pertinentes pour l'analyse lexico-syntaxique constitue l'étape préliminaire d'INSYSE. Il s'agit d'identifier les phrases décrivant les fonctions de gènes dans les maladies du système nerveux : sont sélectionnées celles comportant des lexèmes causatifs tels que *causing, triggering, activating,* etc. Cette étape manuelle est guidée par la liste des relations abstraites de causation du thesaurus Roget. Certains marqueurs de causalité plus spécifiques aux domaine, révélés par l'analyse terminologique de Nomino (Dumas et al. 1997) sur le corpus, peuvent également être utilisés.

L'analyse syntaxique des phrases sélectionnées est basée sur l'utilisation de l'analyseur de surface RASP (Briscoe et Carroll 2002). Pour chaque phrase, les fonctions syntaxiques des lexèmes qu'elle contient sont identifiées et un arbre de dépendance est construit dont les noeuds sont les lexèmes et qui rend compte des fonctions syntaxiques de ces derniers. RASP attribue aux lexèmes d'une phrase les informations lexico-syntaxiques suivantes :

- des relations de dépendance syntaxique ; par exemple, dans le groupe nominal *dialysis patients*, *patients* est la tête et *dialysis* dépend de *patients* ;
- des relations grammaticales : *sujet, objet, auxiliaire,* etc. ;
- des étiquettes morphosyntaxiques (pos-tag) qui indiquent la forme sous laquelle est présent un lexème : par exemple le *temps* et le *mode* d'un verbe.

Considérons par exemple la phrase suivante extraite du corpus de textes étudié : *Hypoxaemia triggered cardiovascular events in dialysis patients.* Son analyse par RASP fournit l'arbre de dépendances suivant :

```
[triggered, vvd]
  [Hypoxaemium, nn1] suj
  [[cardiovascular, jj] [events, nn2] ncmod] dobj
  [[in, ii] [[dialysis, nn1] [patients, nn2] ncmod]] iobj
```

Hypoxaemia est identifié comme étant le sujet de la phrase (suj), *triggered* le verbe (tête), *cardiovascular event* le complément d'objet direct (dobj) et *in dialysis patients* le complément d'objet indirect (iobj). cardiovascular est un modifieur de *event*, dialysis un modifieur de *patient*. *hypoxaemia*, *dialysis*, *events* et *patients* sont des noms communs, singuliers (nn1) ou pluriels (nn2), *triggered* un verbe indicatif passé (vvd), *cardiovascular* un adjectif qualificatif (jj) et *in* une préposition (ii).

Le lexique ainsi établi est raffiné par comparaison des lexèmes qu'il contient avec le résultat d'une extraction automatique des termes du même corpus avec Nomino. La catégorisation lexicale que produit cet extracteur terminologique améliore la pertinence pour le domaine des lexèmes issus de RASP. L'émergence de termes cohérents et pertinents au domaine constitue en effet une étape primordiale dans l'extraction du contenu sémantique des textes (Bourigault et Fabre 2000). Par exemple, à partir de la phrase étudiée ci-dessus, Nomino extrait le terme *dialysis_patient* qui remplacera le lexème *patient* dans le lexique construit avec RASP ; *dialysis_patient* hérite des informations lexico-syntaxiques de *patient* en tant que *tête* du groupe nominal (NP) *dialysis_patient* : relation de dépendance avec *trigger* (argument), relation grammaticale avec *trigger* par le biais de *in* (objet indirect). L'arbre de dépendance construit par RASP est donc transformé comme suit :

```
[triggered, vvd] V
  [Hypoxaemium,NP] suj
  [cardiovascular_events, NP] dobj
  [dialysis_patients, NP] iobj(IN)
```

Ainsi chaque lexème sera constitué d'informations lexicographiques du domaine (plus ou moins normatives), mais surtout d'informations morpho-syntaxiques qui seront prises en compte par les grammaires.

3 Acquisition de schémas conceptuels instanciés

La seconde étape d'INSYSE consiste, à partir des informations lexico-syntaxiques extraites et associées aux lexèmes des phrases du corpus analysées, à acquérir des schémas conceptuels représentant ces phrases. Pour ce faire, nous utilisons le système PATR-II (Shieber 1986) caractérisé par un formalisme d'unification de structures qui permet (1) de faire émerger une structure sémantique complexe et cohérente unique à partir de sous-structures simples et (2) de prendre en compte des problèmes de perspective comme dans la passivation ou la nominalisation.

Pour cela, nous établissons un ensemble de règles de grammaire à partir de l'analyse manuelle d'un sous-ensemble représentatif du lexique extrait lors de l'étape précédente. Basées sur l'unification et la contrainte de traits, ces règles permettent, lorsqu'elles s'appliquent à une phrase (aux informations lexico-syntaxiques extraites), de construire un schéma conceptuel instancié. Elles constituent de fait une interface entre syntaxe et sémantique : elles mettent en correspondance des fonctions syntaxiques telles que *sujet*, *objet*, etc. avec des rôles sémantiques tels que *agent*, *patient*, etc. Considérons par exemple les deux règles suivantes extraites du fichier grammatical que nous avons construit pour l'analyse des phrases de causation de notre corpus extrait de Medline :

Rule {Clause}	Rule {Predication}
(1) S -> NP VP	(1) VP -> V NP PP
(2) <S sem pred> = <VP sem pred>	(2) <VP sem pred> = <V sem>
(3) <VP sem pred postag> = VVD	(3) <VP sem arg1> = <NP>
(4) <S AGT> = <NP>	(4) <NP sem case> = dObj
(5) <NP sem case> = Subj	(5) <VP sem arg2> =<PP>
(6) <S PAT> = <VP sem arg1>	(6) <PP sem case> = iObj
(7) <S SET> =<VP sem arg2>	

TAB 1 – *Exemple de deux règles extraites du fichier grammatical fourni à PATR-II*

La règle *Clause* a trait à la structure principale d'une phrase S composée d'un groupe nominal NP et d'un groupe verbal VP (1). Elle stipule que :

- le prédicat sémantique de S sera celui de VP (2) ;
- si VP est un verbe au mode indicatif passé (3) et NP est sujet (5) alors NP a le rôle thématique d'*agent* AGT de S (4) ;
- l'argument sémantique arg1 de VP a le rôle de *patient* PAT de S (6) et l'argument sémantique arg2 de VP a le rôle de *setting* SET de S (7).

La règle *Predication* a trait à la structure verbale d'un groupe verbal VP composé d'un verbe V, d'un groupe nominal NP et d'un groupe prépositionnel PP (1). Elle stipule que :

- le prédicat sémantique de VP sera la structure sémantique de V (2) ;
- si NP est complément d'objet direct (4) alors l'argument arg1 de VP est NP (3) ;
- si PP est complément d'objet indirect (6) alors l'argument arg2 de VP est PP (5).

Ces deux règles s'appliquent conjointement au fichier lexical de notre exemple et permettent de construire avec PATR-II le schéma conceptuel suivant :

```
[cat: S
    AGT:[cat: NP
         lex:Hypoxaemia                      sem:[case: Subj, pred: HYPOXAEMIA]
    PAT:[cat: NP
         lex: cardiovascular_events sem: [case: dObj, pred: CARDIO-EVENT]]
    SET:[cat: PP
         lex: in_dialysis_patients   sem: [case: iObj, pred: DIALYSIS_PATIENT]]
    sem:[pred: [postag: VVD, pred: TRIGGER]]]
```

Ce schéma stipule que *l'agent* de TRIGGER est HYPOXAEMIA, son *patient* est CARDIO-EVENT et son *setting* est DIALYSIS-PATIENT. Des relations sémantiques sont ainsi établies entre les prédicats sémantiques des lexèmes extraits de la phrase analysée.

De plus, devant une forme passive ou nominale de notre exemple, ce même schéma conceptuel serait donc extrait grâce aux règles appropriées (cinq règles pour la passivation et quatre pour la nominalisation).

4 Annotation de documents à partir de schémas conceptuels

Les schémas conceptuels instanciés obtenus à l'issue de l'étape précédente d'INSYSE constituent des annotations sémantiques des résumés de Medline contenant les phrases analysées ; il s'agit dans cette ultime étape de les traduire dans le standard RDF du web sémantique. Pour ce faire une feuille de style XSLT permet de transformer en RDF les résultats de PATR-II exportés dans la syntaxe XML. L'exemple de schéma conceptuel de la partie précédente est ainsi traduit dans l'annotation RDF suivante :

```
<rdf:RDF
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:gal="http://www.sophia.inria.fr/acacia/galien#">
  <gal:Abstract rdf:about="http://www.sophia.inria.fr/acacia/medline#a324">
      <gal:hasForCausationSchema>
        <gal:CausationSchema rdf:about="http://www.sophia.inria.fr/acacia/caus#c287">
          <gal:agent> <gal:Hypoxaemia/> </gal:agent>
          <gal:patient> <gal:CardioEvent/> </gal:patient>
          <gal:setting> <gal:DialysisPatient/> </gal:setting>
        </gal:CausationSchema>
      </gal:hasForCausationSchema>
  </gal:Abstract>
</rdf:RDF>
```

Nous avons choisi GALEN (Rector et al. 1994) comme ontologie de référence pour la construction de nos annotations. Les concepts qui apparaissent dans les annotations sont soit des concepts de GALEN soit des concepts avec lesquels il s'agit d'enrichir GALEN. Pour cela, nous tentons de mettre en correspondance, les lexèmes qui participent aux schémas conceptuels issus de PATR-II avec les concepts existants de GALEN. Cette tâche encore à l'étude repose sur des techniques déjà existantes sur le calcul de similarité sémantiques (Maedche et Staab 2001).

5 Comparaison d'INSYSE avec d'autres approches

La figure ci-dessous synthétise les étapes successives de la méthode INSYSE. Elle est caractérisée par une forte granularité de l'analyse syntaxique (étape 2), et par l'interfaçage syntaxico-sémantique qu'elle opère (étape 3). Elle est en effet basée sur une approche

cognitive fonctionnaliste d'analyse (Nuyts 1992) centrée sur la mise en correspondance de fonctions syntaxiques et de relations sémantiques, reposant sur une correspondance prototypique (forme active), qui est dynamique en fonction de la perspective (forme passive, nominale, dative, *etc*.). INSYSE s'apparente en cela aux méthodes par *patrons lexico-syntaxiques* (pattern matching) dans lesquelles des concepts sont déduits à partir de marqueurs du domaine et par l'analyse de leurs contextes ; COATIS (Garcia 1997) procède ainsi pour extraire des relations de causalité.

INSYSE s'apparente également à ASIUM (Faure et Nédellec 1998) et à OntoLT (Buitelaar et al. 2004) qui soulèvent l'importance des fonctions grammaticales pour comprendre les liens entre prédicat et arguments pour l'interprétation conceptuelle. Cependant ces deux approches opèrent un *pattern matching* direct entre fonctions grammaticales et *domain* et *range* des propriétés par l'intermédiaire de règles de correspondance prédéfinies; de plus ASIUM adopte une approche statistique du traitement des informations lexico-syntaxiques.

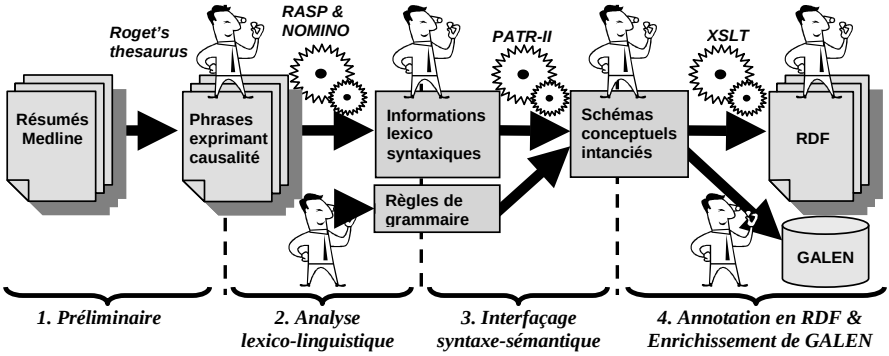

FIG. 1 – *Les étapes majeures d'INSYSE*

INSYSE se caractérise également par une analyse des seules relations sémantiques de *causation*, l'annotation des résumés médicaux étant guidée par un but: celui d'expliciter les fonctions des gènes et leurs relations de causalité avec les maladies du système nerveux, même si ce type de corrélation peut être exprimé par un type de relation autre que la causalité. En outre, la finalité d'INSYSE est la construction d'annotations exprimant ces relations de causalité; l'apprentissage de concepts est corollaire au processus d'annotation.

Enfin, INSYSE est une méthode *semi-automatique* d'extraction des connaissances et s'apparente en cela à l'approche décrite dans (Aussenac-Gilles et al., 2003) dont les auteurs plaident en faveur d'une forte implication des experts pour valider les connaissances, outre l'utilisation d'outils de traitement automatique de la langue naturelle.

6 Conclusion et Perspectives

INSYSE est une méthode semi-automatique d'annotation de textes que nous appliquons à un corpus de résumés médicaux pour lesquels il s'agit d'expliciter des relations de causalité entre certaines maladies du système nerveux et les fonctions des gènes. INSYSE est centrée sur l'extraction d'informations lexico-syntaxiques à partir de phrases comportant des

lexèmes de causation et sur l'acquisition de schémas conceptuels instanciés à partir de ces informations, grâce à un ensemble de règles de grammaires d'unification dédiées.

Nous avons analysé un corpus de 5000 résumés, pour lequel nous avons identifié environ 40 schémas de causation non instanciés. Le fichier grammatical doit encore être enrichi et l'élaboration d'une feuille de style XSLT pour automatiser la traduction des schémas en RDF est en cours. C'est à travers cette écriture que les schémas conceptuels *non instanciés* deviennent explicites. Nous projetons de les exploiter pour définir des requêtes RDF(S) permettant avec Corese d'interroger le corpus de résumés annotés. Nous les envisageons également en tant qu'axiomes du domaine qui permettraient d'enrichir l'ontologie.

Références

Aussenac-Gilles N., Biebow B., et Sulzman S. (2003), D'une méthode à un guide pratique de modélisation de connaissances à partir de textes. Actes des 5èmes rencontres TIA, 2003.

Bourigault D. et Fabre C. (2000), Approche linguistique pour l'analyse syntaxique de corpus, Cahiers de Grammaires, **25**, 131-151, 2000.

Briscoe T. et Carroll J. (2002), Robust accurate statistical annotation of general text, Proceedings of the 3rd International Conference on Language Resources and Evaluation, pp 1499-1504, 2002.

Buitelaar P., Olejnik D., et Sintek M. (2004), A Protege plug-in for ontology extraction from text based on linguistic analysis, ESWS'04, Heraklion, 2004.

Corby O., Dieng-Kuntz R. et Faron-Zucker C. (2004), Querying the semantic web with the Corese search engine, ECAI'2004, Valencia, August 2004, IOS Press, p. 705-709.

Dumas L., Plante A. et Plante P. (1997), ALN : Analyseur Linguistique de ALN, ATO, 1997.

Faure D. et Nédellec C. (1998), A corpus-based conceptual clustering method for verb frames and ontology acquisition, Proceedings of LREC workshop on Adapting lexical and corpus resources to sublanguages and applications, pp 5-12, 1998.

Garcia D. (1997), COATIS: a NLP system to locate expressions of actions connected by causality links, Proceedings of EKAW'97, pp 347-352, 1997.

Maedche A., et Staab S. (2001), Comparing Ontologies: Similarity Measures and a Comparison Study. Internal Report 408, Institute AIFB, University of Karlsruhe, 2001.

Nuyts J. (1992), Aspects of a Cognitive-Pragmatic Theory of Language, Benjamins, 1992.

Rector A., Gangemi A., Galeazzi E., Glowinski A. et Rossi-Mori A. (1994), The GALEN Model Schemata for Anatomy, MIE'94,. Lisbon, 1994.

Shieber S.M. (1986), An Introduction to Unification-Based Approaches to Grammar, University of Chicago Press, Chicago (CSLI Lecture Notes Series, 4), 1986.

Summary

We present the INSYSE method (Interface Syntax Semantics) for annotation of texts, based on extraction of semantic relations from syntactic structures. We apply this method to a corpus of 5000 Medline abstracts about central nervous system diseases and genes interactions. Our aim is to build semantic annotations for these abstracts so as to query the corpus about functions of genes and their correlations with particular diseases. Our semi-automatic approach focuses on (1) extracting lexico-syntactic information from sentences in the corpus comprising causation lexemes and (2) elaborating unification grammar rules which enable to extract instantiated conceptual schemas from this information. They are translated into RDF annotations which are used by the semantic search engine Corese to query the corpus.