Extraction of Lexico-Syntactic Information and Acquisition of Causality Schemas for Text Annotation

Laurent Alamarguy[1], Rose Dieng-Kuntz[1], and Catherine Faron-Zucker[2]

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

Abstract. We present the INSYSE method for the 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 gene interactions. Our cooperative 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 used by the semantic search engine Corese to query the corpus about functions of genes and their correlations with particular diseases.

1 Introduction

The notion of causality is essential to understand some correlations in functional genomics. The automation of the detection of such causality correlations and their conceptual representation is a keystone to build a community memory. This can be achieved by using some Natural Language Processing (NLP) methods.

We propose a semi-automatic method of text annotation which is based on the acquisition of conceptual templates from the extraction of lexico-syntactic structures. We call it INSYSE (Interface of SYntax-SEmantics). It is applied to a corpus about 5000 biomedical abstracts from Medline, dealing with central nervous system pathologies and the gene interactions in these pathologies. We aim at generating semantic annotations on these abstracts to inform about gene functions and their causal relations with some diseases. A memory of the community of the actors in biomedical field can thus be built.

INSYSE only focuses on causation relation analysis, since the aim of detecting some correlation between gene functions and pathologies favors this focus and our corpus is characterized by numerous and various causation markers. However, some other relationships certainly underlie in the comprehension of these correlation, but we do not address their study in this work. We study *intra-clausal* causation markers; discourse markers, that may overlap several sentences, are out of the scope of our study, since their construal and processing require another linguistic analysis.

R. Khosla et al. (Eds.): KES 2005, LNAI 3683, pp. 1180–1186, 2005.

In this paper, we introduce the various steps of the INSYSE method, as depicted in Figure 1. INSYSE stresses on the processing of a fine grained syntactic analysis (step 2), and the construal of an accurate syntax-semantic interface (step 3). The second stage mainly relies on the merging of a terminological extraction with a partial syntactic parsing, so as to provide domain-relevant concepts and accurate interconnections between these concepts. The syntax-semantic interface is based on a cognitive-functionalist approach [10] advocating a strong correlation between semantic roles and syntactic functions from prototypical mapping (active form) and from dynamic operations such as *perspectivization* enabling to construe passive or nominal form, or dative shift.

In section 2 we describe the extraction step of lexico-syntactic information through sentences containing some causation lexemes. Section 3 is dedicated to the elaboration of rules based on unification grammars which enable to extract some lexico-syntactic information peculiar to some instantiated conceptual schemas. In section 4, we describe how these schemas are translated into RDF(S)[1] annotations from which the corpus will be queried through the inference search engine CORESE [5], once a concept matching phase will have been processed. In conclusion, the INSYSE method is compared with other approaches related with text annotations and we sketch its on-going evaluation.

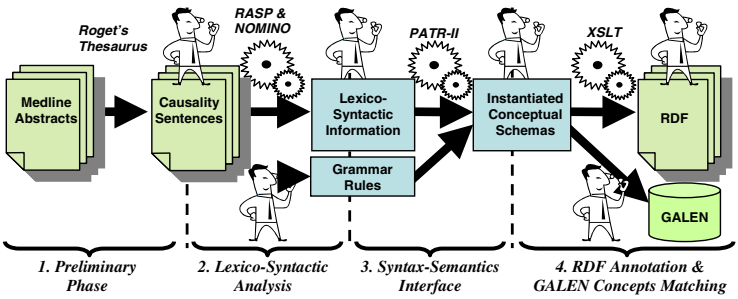


Fig. 1. INSYSE in a Nutshell

2 Lexeme Extraction from Texts for a Lexicon Construction

The INSYSE preliminary step consists in the selection of relevant sentences, from abstracts in our corpus, so as to operate the lexico-syntactic analysis. It aims at identifying the sentences describing gene functions interacting in nervous system pathologies, and the relevant sentences are selected according to the causative lexemes they contain, such as *causing*, *triggering*, *activating*, etc. This stage is guided by the abstract relations of causation listed in the Roget's Thesaurus.

The syntactic analysis of the selected sentences is based on the application of the RASP shallow parser [3] on the whole corpus. So for each sentence, the

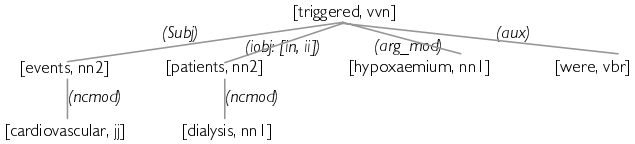
[1] http://www.w3.org/TR/1999/REC-rdf-syntax-19990222

syntactic functions of its lexemes are revealed. A dependency tree is built with a lexeme for each node and drawn from syntactic functions. RASP assigns to each lexeme the following lexico-syntactic information:

- syntactic dependency relations, *e.g.*, in NP *dialysis patients*, the noun *patients* symbolizes the 'head' and *dialysis* is the dependency of *patients*;
- grammatical relations such as *subject*, *object*, *auxiliary*, etc.;
- morphosyntactic tags (PoS tags) indicating the grammatical category of each word through context.

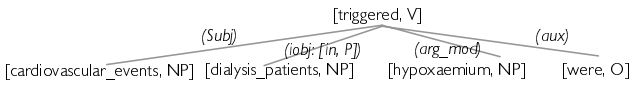
Let us consider the following excerpt of our training corpus:

Cardiovascular events were triggered in dialysis patients by hypoxaemia. The parsing of RASP construes the following dependency tree:



Events is construed as the sentence subject (*subj*), *triggered* as the verbal predicate (*head*), *patients* as the indirect object, specified by the preposition *in* (*iobj: in*), *hypoxaemia* as an adjunct (*arg_mod*), and *were* as the auxiliary (*aux*). *Cardiovascular* and *dialysis* are construed as modifier of respectively *events* and *patients*. *Hypoxaemia*, *dialysis*, *events* and *patients* are commoun nouns, singular (*nn1*) or plural (*nn2*), *triggered* is a past-participle verb (*vvn*), *were* is a preterit form of the *be* auxiliary (*vbr*), *cardiovascular* is a general adjective (*jj*) and *in* a preposition (*ii*).

The constructed lexicon is refined by comparing the lexical entry embodied by each lexeme with the automatic term extraction operated by Nomino [6] on the same corpus. This terminological extractor provides a more accurate syntagmatic categorisation enhancing the relevance of lexical entry to the domain vocabulary. So the revelation of coherent and relevant domain terms constitutes a fundamental step in semantic extraction from texts [2]. For instance, Nomino analysis extracts from the above sentence the term *dialysis_patient*; it will replace the RASP lexeme *patient* as a lexical entry. *Dialysis_patient* inherits the lexico-syntactic information of *patient*, which is the head of the nominal phrase (NP) *dialysis patients*: its dependency relation with *triggered* as an argument, its grammatical relation with *triggered* introduced by *in* (indirect object). The RASP dependency tree then becomes:



Thus, each lexical entry is constituted of lexicographical domain information, and morpho-syntactic information that will be processed by our grammar rules.

3 Instantiated Conceptual Schemas Acquisition

The second stage of INSYSE consists of acquiring conceptual schemas capturing the meaning of a sentence, from the lexico-syntactic information associated with each lexical entry extracted from parsed corpus sentences. To achieve it, we use the grammatical parser PATR-II [12] defined by a unification formalism, and enabling (1) to reveal a peculiar complex and coherent semantic structure from more primitive substructures, and (2) to construe *perspective* grammatical operations such as passivation or nominalization.

So, we have defined a set of about 50 grammar rules from the manual study of representative causation constructions in the training corpus. Based on feature unification and constraints, rules parse a sentence using the extracted lexico-syntactic information and build an instantiated conceptual schema. Thus, these rules embody the syntax-semantics interface, since they map syntactic functions such as *subject*, *object*, etc. with semantic functions like *agent*, *patient*, etc. The following five rules in Table 1, extracted from the grammar we have built and dedicated to the causation construal, parse passivation:

Table 1. Example of grammar rules processed by PATR-II

```
   Rule {Clause Passivation}
(1) S -> NP VP
(2)    <S sem pred> = <VP sem pred>
(3)    <VP sem postag> = VVN
(4)    <S AGT> = <VP sem arg2>
(5)    <S AGT sem case> = Arg_Mod
(6)    <S PAT> = <NP>
(7)    <NP sem case> = Subj
(8)    <S SET> =<VP sem arg1>

   Rule {Passive Predication}
(1) VP -> V2 PP1 PP2
(2)    <VP sem pred> = <V2>
(3)    <VP sem arg1> = <PP1 sem pred>
(4)    <VP sem arg2> = <PP2 sem pred>

   Rule {Passive Predication Operator }
(1) V2 -> O V
(2)     <V2 sem pred> = <V sem pred>
(3)     <V2 postag> =<V sem postag>
(4)     <V2 sem arg> = <O>
(5)     <O sem case> = Aux

   Rule {Periphery1}
(1) PP1 -> P NP
(2)     <PP1 sem pred> = <NP>
(3)     <PP1 sem arg> = <P>

   Rule {Periphery2}
(1) PP2 -> P NP
(2)     <PP2 sem pred> = <NP>
(3)     <PP2 sem arg> = <P>
```

The rule *Clause Passivation* refers to the passive form of a sentence S, constituted with a noun phrase NP and a verb phrase VP (1), and stipulating that:

- The semantic predicate of S will be inherited from VP (2);
- if VP is a past participle verb (3), NP is subject (7) and the *agent* role AGT is filled by the adjunct Arg_Mod (5), then NP fulfills the *patient* semantic role PAT of S (6), and the semantic argument arg2 of VP plays the AGT role of S (4);
- the semantic argument arg1 of VP plays the *setting* role SET of S (8).

The rule *Passive Predication* refers to the passive form of a VP predicative structure, constituted with a verbal structure V2, and two prepositional phrases PP1 and PP2 (1), stipulating that:

- the semantic predicate of VP will be inherited from V2 (2);
- the argument arg1 of VP will be inherited from the semantic predicate of PP1 (3);
- the argument arg2 of VP will be inherited from the semantic predicate of PP2 (4);

The rule *PassivePredication Operator* refers to the nucleus structure of a verbal constituent V2, constituted with an operator O and a verb V (1), and stipulating that:

- the semantic predicate of V2 will be inherited from the semantic predicate of V (2);
- the morpho-syntactic category of V2 will inherit the morpho-syntactic tag of the semantic structure of V (3);
- if the operator O is auxiliary (5), then O becomes argument of V2 (4).

The rules *Periphery1* and *Periphery2* refer to a PP prepositional structure, constituted with a preposition P and a noun phrase NP (1), both stipulating that:

- the semantic predicates of both PP1 and PP2 correspond to the nominal phrase NP (2), and the arguments of PP1 and PP2 correspond to the preposition P (3).

When processing the lexicon file extracted from the above sentence taken as example, these five rules parse the following conceptual schema through PATR-II:

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

This schema stipulates that the *agent* of TRIGGER is fulfilled by HYPOXAEMIA, the *patient* of TRIGGER is fulfilled by CARDIO-EVENT and the *setting* of TRIGGER is fulfilled by DIALYSIS-PATIENT. Thus, semantic relations interconnect semantic predicates of lexemes extracted from the corpus.

Moreover, this conceptual schema would be also elaborated to construe an active or nominalized form, following the relevant rules.

4 Document Annotation from Conceptual Schemas

These acquired instantiated conceptual schemas will constitute semantic annotations of the Medline abstracts whose sentences have been parsed. This last stage aims at translating these schemas into the RDF semantic web standard language. The output of the PATR-II parsing in XML syntax is converted into RDF by using a XSLT style sheet. For instance, the above conceptual schema is translated into the following RDF annotation:

```
<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:sem> <gal:trigger/> </gal:sem>
         </gal:CausationSchema>
       </gal:hasForCausationSchema>
  </gal:Abstract>
</rdf:RDF>
```

It is worth noticing that the RDF annotation solely keeps semantic information from the instantiated conceptual schema, and is pruned from all syntactic features. Furthermore, a validation analysis on these semantic annotations is elaborated by domain experts that only retain accurate and relevant ones.

5 Related Work

INSYSE is close to pattern matching methods, that deduce concepts from domain semantic markers and through their contextual analysis; COATIS [8] adopts this approach to extract causality relations. INSYSE is also close to ASIUM [7] and OntoLT [4] that stress the importance of grammatical relations to apprehend the interconnections between concepts. However, these approaches perform a direct pattern matching between syntactic parsing and semantic annotation, without an intermediary fine grained semantic construal. Moreover, ASIUM syntactic information process relies on statistics. INSYSE is a semi-automatic knowledge extraction method, close to the approach proposed in [1].

6 Conclusion and Perspectives

We have presented INSYSE, a semi-automatic text annotation method applied in biomedical domain, aiming at construing causation relations implying genes functions in central nervous system pathologies. INSYSE focuses on the acquisition of causation instantiated conceptual schemas, construed by a set of dedicated unification grammar rules processing a lexicon based on the merging of a terminological extraction with a partial syntactic analysis.

The main contribution of our paper is twofold: first we advocate the processing of a fine grained syntactic analysis, by merging a terminological processing with a shallow syntactic parsing; secondly we favour an accurate syntax-semantic interface through a fine grained semantic construal operated by grammar rules and processed by PATR-II grammatical parser.

A first implementation of INSYSE in Java has just been carried out. We are currently making some adjustments to apply our system to the analysis of the whole corpus of 5000 Medline abstracts. From a linguistic viewpoint, we want to evaluate the accuracy of grammar rules we have built together with the whole linguistic process, by analysing the annotation generation – or none generation

– expected for each causality sentences. From the genomic domain viewpoint, experts should validate the relevancy of the semantic annotations iteratively generated.

As further work, versioning and backward interaction between the stages of INSYSE would be useful for validation and adjustment purposes. Second, even if the causality sentence identification in stage 1 is not the core of our work, we can fairly enhance it with more domain-specific causality markers, and those revealed by the terminological analysis of Nomino [6] on the corpus, may also be useful for this task. Finally, the finalization of our annotation construction will be effective with an ontology concept matching stage, so as to obtain consensual semantic annotations. This stage aims at mapping each term filling our PATR-II conceptual schemas with some GALEN [11] concepts. Two different approaches are currently tested for this mapping task: the first one relies on a lexicographic similarity calculus, based on tokens or lemmas analysis; the second one relies on an ontology integration based method, by using semantic similarity calculus described in [13].

References

1. Aussenac-Gilles, N., Biebow, B., Sulzman, S.: Revisiting Ontology Design: a methodology based on corpus analysis. In *Proceedings of EKAW'2000* (2000) 172-188
2. Bourigault, D., Fabre, C. : Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires 25* (2000) 131-151
3. Briscoe, T., Carroll, J.: Robust accurate statistical annotation of general text. In *Proceedings of LREC'02* (2002) 1499-1504
4. Buitelaar, P., Olejnik, D., Sintek, M.: A Protege plug-in for ontology extraction from text based on linguistic analysis. In *Proceedings of ESWS'04* (2004)
5. Corby, O., Dieng-Kuntz, R., Faron-Zucker, C.: Querying the semantic web with the Corese search engine. In *Proceedings of ECAI'2004* (2004) 705-709.
6. Dumas, L., Plante, A., Plante, P. : *ALN : Analyseur Linguistique de ALN*. ATO, UQAM (1997)
7. Faure, D., Nédellec, C.: A corpus-based conceptual clustering method for verb frames and ontology acquisition. In *Proceedings of LREC workshop on Adapting lexical and corpus resources to sublanguages and applications* (1998) 5-12
8. Garcia, D.: COATIS: a NLP system to locate expressions of actions connected by causality links. In *Proceedings of EKAW'97* (1997) 347-352
9. Maedche, A., Staab, S.: *Comparing Ontologies: Similarity Measures and a Comparison Study*. Internal Report, University of Karlsruhe (2001)
10. Nuyts, J.: *Aspects of a Cognitive-Pragmatic Theory of Language*. Benjamins (1992)
11. Rector, A., Gangemi, A., Galeazzi, E., Glowinski, A., Rossi-Mori, A.: The GALEN Model Schemata for Anatomy. In *Proceedings of MIE'94* (1994)
12. Shieber, S.M.: *An Introduction to Unification-Based Approaches to Grammar*. CSLI Lecture Notes Series, vol. 4. University of Chicago Press, Chicago (1986)
13. Wang, H., Azuaje, F., Bodenreider, O., Dopazo, J.: Gene Expression Correlation and Gene Ontology-Based Similarity: An Assessment of Quantitative Relationships. In *Proceedings of CIBCB'04* (2004) 25-31