PredXtract, a generic platform to extract in texts predicate argument structures

Elisabeth Godbert, Jean Royaute

PredXtract, a generic platform to extract in texts predicate argument structures (PAS)

Elisabeth Godbert, Jean Royauté

Laboratoire d'Informatique Fondamentale de Marseille (LIF)
CNRS UMR 6166 - Université de la Méditerranée
Parc Scientifique et Technologique de Luminy, case 901
13288 Marseille Cedex 9
Elisabeth.Godbert@lif.univ-mrs.fr, Jean.Royaute@lif.univ-mrs.fr

Abstract

Verbal and nominal predicate structures present interesting properties for information extraction. We show how to study these predicate structures in a uniform way, using the fact that the nominalization of a verb has the same arguments as the verb. We then describe the extraction platform (PredXtract) which we have developed in order to extract predicate argument structures and which highlights relations between biological entities in biological texts. We present and discuss our results.

1. Introduction

This paper focuses on the extraction of verbal and nominal predicate structures, which can be expressed in a great variety of forms (Meyers et al., 2004a). Defining a uniform representation for these structures is decisive to converge on a VerbNet or FrameNet representation (Wattarujeekrit et al., 2004; Miyao et al., 2006; Levin, 1993) and to acquire semantic relations.

In predicate-argument representation, verbs and their nominalizations are the most productive predicates and have the same argument relations, where arguments play precise conceptual roles: subjects and complements, which are core arguments, and adjuncts. With a nominalization, it is possible to build complex noun-phrases (NPs), in which the head noun is bound to prepositional phrases (PPs) with specific prepositions which mark core arguments or adjuncts. For example, the NP *milk concentration by ultrafiltration* is related to the sentences *ultrafiltration concentrates milk* and *milk is concentrated by ultrafiltration*: the NP is built with the predicate head *concentration*, preceded or followed with its arguments *ultrafiltration* and *milk*, whether or not it is introduced by a preposition. In these structures, the core arguments are preserved and it is possible to insert an adjunct (*in the manufacture of cheese*).

Verbal and nominal structures are closely correlated; we will show in Section 2. how to link an NP built with a nominalization, to a core sentence. We use the following notation: N_0 *V W*, where N_0 is the suject of the verb, V the verb and W, a sequence of complements (N_1 ... N_n) linked to the verb (Gross, 1986).

Our objective is to define a underspecified semantic representation where (i) the predicate (nominal or verbal) expresses the action, (ii) the subject is the Agent (who performs the action) (iii) the object is the Patient (who is involved by the action) and (iv) possible adjuncts express context of the action ; this semantic role is named here Circumstance. Thus, this representation is situated at the syntax-semantics interface. Distinguishing the core arguments from the adjunct arguments in predicate structures (Tesnière, 1959) is important in information extraction and particularly in scientific sublanguage. Later, this semantic representation will be enriched by including more complex roles derived, for example, from semantic frames of VerbNet.

At present, we have developed a robust platform, PredXtract, based on the Link Parser (Sleator and Temperley, 1991). This platform is a generic tool which extracts verbal and nominal predicate argument structures (PAS) in English texts. More specifically, it exhibits relations between biological entities.

2. Nominal and verbal argument structures

We present here a typology of seven classses of verbal and nominal structures, defined from their core arguments:

- Verbs accepting a direct object are grouped together in Class 1 and 2; in the corresponding predicate noun phrases (PNPs), the preposition *of* marks the direct object.
- Verbs that do not accept a direct object are grouped together in Class 3 to 5; in the corresponding predicate noun phrases (PNPs), the preposition *of* marks the subject.
- Symmetric predicates with interchangeable arguments concern Class 6 and 7.

This classification has been elaborated, from scientific texts of the web, and from the grammar of English described in (Quirk et al., 1987), as well as from the data of "The Specialist Lexicon", which gives, for all verbs, their nominalizations and the different prepositions that can introduce core arguments (www.nlm.nih.gov/pubs/factsheets/umlslex.html).

2.1. The preposition *of* as marker of the object

Class 1: N_0 *V* N_1 = N^{pred} *of* N_1 *by* N_0. This class groups together predicates with a direct objet and which accept passive voice (N_1 *is* V_{ed} *by* N_0). This is the most important class with more than 1,000 couples of verbs/nominalizations. For example, the couple *activate/activation* belongs to this class : *IFN-gamma activates protein kinase C delta / activation of protein kinase C delta by IFN-gamma*.

Class 2: N_0 *V* N_1 *Prep* N_2 = N^{pred} *of* N_1 *Prep* N_2 *by* N_0. This class concerns constructions with a direct object and with a second complement introduced with a preposition inherited from the verbal construction. This preposition is the same in the verbal construction and the nominal construction. These constructions also accept passive voice. Example: N_0 *attributes a protein fragment* ***to*** *a sequence / attribution of a protein fragment* ***to*** *a sequence by* N_0.

2.2. The preposition *of* as marker of the subject

Class 3: N_0 *V* = N^{pred} *of* N_0. This class concerns constructions without complement. In the NP construction, the preposition *of* introduces the subject argument. Example: *the femoral head necroses / necrosis of the femoral head*.

Class 4: N_0 *V Prep* N_1 = N^{pred} *of* N_0 *Prep*N_1. This construction can appear without a prepositional complement but if the complement is present, the same preposition introduces it in the sentence and in the NP. As in Class 3, the preposition *of* marks the subject argument in the NP. Example: *tryptophans fluctuates in gramicidin / fluctuation of tryptophans in gramicidin*.

Class 5: N_0 *V Prep* N_1 *Prep* N_2 = N^{pred} *of* N_0 *Prep* N_1 *Prep* N_2. In this class, the two prepositions which appear in the sentence also appear in the NP. Example: *temperature decreases from 200 K to 70 K / decrease of temperature from 200 K to 70 K*.

2.3. Predicates with permutable arguments

Class 6: N_a *V with* N_b = N^{pred} *of* N_a *with* N_b = N^{pred} *of/between* N_a *and* N_b. This is a special class because the arguments can permutate without a change in the meaning. For that reason we noted them N_a and N_b.
Examples: *genes interact with proteins*; *interaction of genes with proteins / interaction of/between genes and proteins*.

Class 7: N_0 *V* N_a *Prep* N_b = N^{pred} *of* N_a *with/to* N_b *by* N_0 = N^{pred} *of/between* N_a *and* N_b *by* N_0. We consider that this class is a variant of Class 6 because N_a and N_b are in the complement position in the sentence. For example, from the sentence N_0 *connects a new sequence with/to a cluster*, it is possible to derive several NPs : *connection of a new sequence with/to a cluster / connection of/between a new sequence and a cluster*. In these different constructions, the N_0 argument can be absent in the sentence or in the NP.

In all classes, the arguments introduced by prepositions *of* or *by* can be in the position of left modifier of the nominalization (*regulation of VEGF by TGFbeta1 / VEGF regulation by TGFbeta1 / TGFbeta1 Regulation of VEGF*).

3. PredXtract, an extracting platform

The PredXtract platform produces the representation of a sentence in a set of complex predicate argument structures. PredXtract uses the Link Parser (LP) and its English native Link Grammar (LG), a variant of dependency grammars (Sleator and Temperley, 1991). The sentence processing of the LP produces a set of graphs where words are linked in pairs by labeled arcs with grammatical functions; each graph corresponds to a possible analysis. In LG, generic links attach verbs (`MVp` link) or nouns (`Mp` link) to any preposition which introduces an NP.
In order to mark the precise role of each argument of the predicates, we have: (i) defined specific argument links, in order to distinguish core arguments from adjunct arguments during the extraction process; (ii) integrated in the native grammar of the LP, a grammatical module to parse predicate NPs with specific argument links; (iii) post-processed the parse to align argument links of the verbs to the argument links defined for nominalizations; (iv) modified the classification heuristics of the LP parses because they are not always adapted to biomedical texts (Pyysalo et al., 2006) and because the predicate NP attachments are often not correct.

Besides, to enhance the accuracy of the parsing, we have followed Szolovits (2003) and added in the grammar all of the words of "The Specialist Lexicon" (SL), which includes UMLS terms. We have also added a lexicon of genes and proteins extracted from corpus. The lexicon contains about 400,000 lexical items (500,000 inflected forms).

We describe below the different processes and compoments of PredXtract.

Link Grammar of nominalizations.
Several teams in biomedecine use the LP but without modifying its grammar (Ding et al., 2003; Hakenberg et al., 2009). This parser is also used in other domains as the information extraction in Reuter corpus (Madhyastha et al., 2003). According to our classification of the nominalizations,
we have added to the native LP a grammar module of PNPs in which about 3,900 nominalizations are divided into 89 subclasses. Each subclass corresponds to a syntactic pattern with core arguments (including clauses with *that*) and adjuncts.
All of the words in the LG appear in the same format: just the inflected form or the inflected form followed by a dot and an extension. The extension (a short sequence of alphanumeric characters) allows to re-use the same word in different disjoint linguistic descriptions. Each nominalization belongs to one or more subclasses and can accept one or more syntactic descriptions; in these cases, specific extensions are used.
Figure 1 shows several examples of extensions: the `nt0` extension corresponds to the nominalizations of transitive verbs (*regulate / regulation, product / production, accumulate / accumulation*), `ni2` (*respond / response*) corresponds to the nominalizations of prepositional verbs with the preposition *to*, and `ndt7` (*treat / treatment*) corresponds to the nominalization of the verb with a direct object and a complement introduced by the *with* preposition.
We see in Figure 1 parses of two short sentences with five nominalizations. In the first sentence (example 1), *response* has two arguments : the `MSI` link marks the subject introduced by the preposition *of*, while the `MCITO` link marks the complement introduced by *to*. The second NP shows the prepositional use of *treatment*, not saturated in this case:

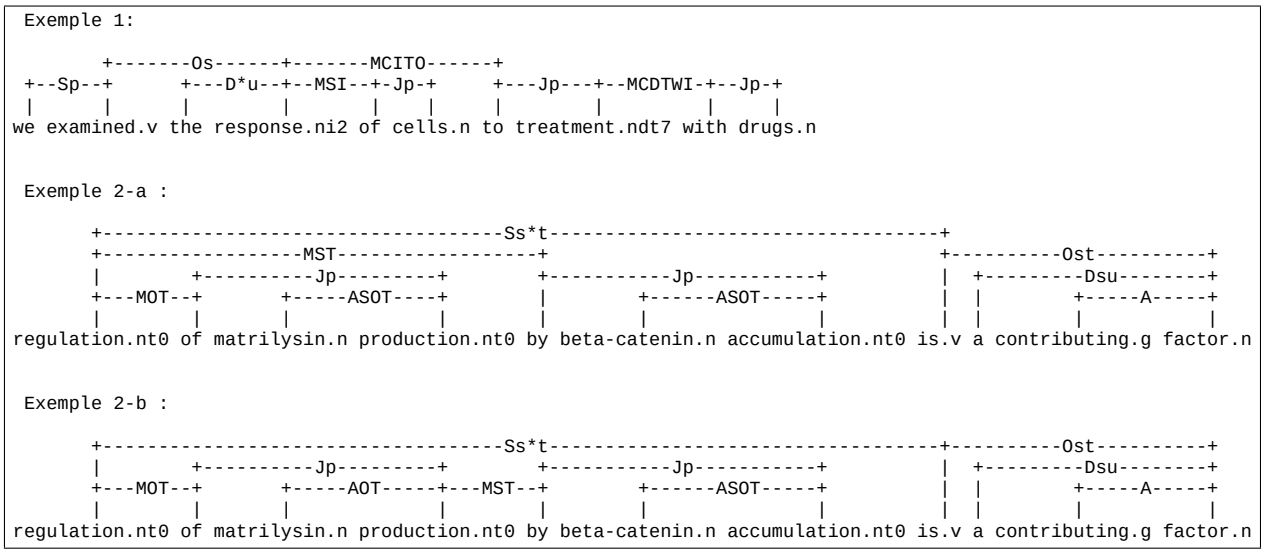

Figure 1: LP parses with several nominalizations.

it has only one argument, introduced by the preposition *with* (link MCDTWI) inherited from the verb.
In Example 2, the two nominalizations *production* and *accumulation* have a left argument marked by the ASOT link. This link means that the argument can be subject or object. In this case, the argument role remains underspecified in this modifier position, because it is not possible to specify the argument role when a prepositional position is lacking.

Verb-noun alignment.
The native grammar of the LP does not distinguish core arguments and adjunct arguments for verbs: it marks all prepositional complements with the same link (MVp).
Rather than writing a grammar for verbs, which would have been very complex, we have defined a module that aligns verb arguments to nominalization arguments during a post-processing step. This module therefore produces a representation of verbs similar to the representation of nominalizations. For this, we use the data of "The Specialist Lexicon" (SL) which gives, for all verbs, the prepositions that can introduce a core argument. This module performs several tasks: (i) distinguish complements from adjuncts of verbs, by using the data of SL, and substitute the generic MVp link with a specific argument link when appropriate; (ii) identify each "verbal sequence" (compound with a verb and a set of possible auxiliaries, negation, and modal verbs); (iii) identify arguments in passive or active voice, and interchangeable arguments.

Recognition of syntactic arguments.
For each parse of a sentence, all of the predicates and their arguments are identified. Each argument link points on the head of a core argument or on a word which introduces it (a preposition or a conjunction). Then the surface structure of each argument is reconstructed via the links, by using linguistic criteria. The reconstructed arguments can be NPs (most cases), clauses or adverbs.

Filtering of parses.
For each sentence, the parses (often several thousands) are re-ordered by attributing to each parse a score defined through several criteria. Among the main criteria:
(i) in the case of multiple prepositional attachments to verbs or nouns, we favor parses whose number of argument links is maximum - a higher score is given to these parses;
(ii) for the treatment of PNPs containing several nominalizations, we favor prepositional arguments attached to the head of the PNP; a specific score is calculated in the case of these PNPs.
This second point is illustrated in (Figure 1, examples 2-a and 2-b) with the two parses of the same sentence. In this sentence, with three nominalizations derived from transitive verbs, the preposition *by* can be attached either to *regulation* or to *production* with the MST argument link. We favor the parse given in example 2-a, because the first nominalization (*regulation*) is in a saturated form, ie. with all core arguments: the subject argument (*matrilysin production*) is marked with MST link and the object argument (*beta-catenin accumulation*) is marked with MOT link.

Syntax-semantics interface.
PredXtract produces for each sentence an underspecified semantic representation, which is close to the syntax. As we have seen, we separate core arguments from adjunct arguments. On this basis, we identify core arguments in several alternation forms. According to Cohen et al. (2008), we extend the paradigm of alternations to the nominalizations. For example, sentences with the different variants of the verb *regulate* : (i) *Fatty acids and eicosanoids regulate gene expression*; (ii) *telomerase activity is mainly regulated by hTERT*; (iii) *A DNA binding protein regulated by IL-4*; (iv) *A unique mechanism regulating gene expression*; (v) *the regulation of eIF4E by 4E-BP phosphorylation is performed at its free state*; (vi) *this study reports the first evidence of VEGF regulation by heregulin in cancer cells* illustrate the different surface forms of similar PAS. To unify the semantic representation, two macro-roles have been defined: Agent and Patient. These two macro-roles are present in these examples with the same PAS. Sentence (i) is in the active voice and the two NPs *Fatty acids* and

eicosanoids are the Agent and *gene expression* the Patient; sentence (ii) is in the passive voice: the Agent is *hTERT* and *telomerase activity* is the Patient; sentences (iii) and (iv) show this verb in participial modifier forms (past and present) with respectively *IL-4* and *A unique mechanism* as Agent and *A DNA binding protein* and *gene expression* as Patient; finally, the last sentences (v) and (vi) show two nominal forms of *regulate* with respectively *4E-BP phosphorylation* and *heregulin* as Agent and *eIF4E* and *VEGF* as Patient.

At present, PredXtract does not take into account all possible syntactic alternations, which correspond to the different ways in which verbs can express their arguments (Levin, 1993).

4. Results and discussion

4.1. PredXtract outputs

In this section, we present results obtained with PredXtract, by showing, for each sentence, the analysis which has obtained the best score. In the first and the two last examples, the extraction of all the predicates and their arguments were identified, and the analysis is correct. In the two others, almost all the predicates and their arguments were identified, but the analysis is not completely correct.

Example 1: from the sentence *Hyperoxic exposure induced an S-phase arrest associated with acute inhibition of Cdk2 activity and DNA synthesis*, 9,168 parses were found and PredXtract outputs:

```
------------------------------------------------------
Nominalization 1: exposure

Nominalization 2: arrest
 Agent or Patient: S-phase

Nominalization 3: inhibition
 Patient: Cdk2 activity
 Patient: DNA synthesis

Nominalization 4: synthesis
 Agent or Patient: DNA

Verb 1: induced (verbal sequence: induced ; active)
 Agent: hyperoxic exposure
 Patient:  an S-phase arrest associated
    with acute inhibition of [...] synthesis

Verb 2: associated (verbal sequence:
                              associated ; passive)
 Patient A: an S-phase arrest
 Patient B: acute inhibition of Cdk2 activity
    and DNA synthesis
------------------------------------------------------
```

This example shows a short sentence with six predicate structures. We can notice that (i) *exposure* has no argument, (ii) *inhibition* has two coordinated Patient roles, (iii) the role of the argument of *arrest* and *synthesis* is underspecified (Agent or Patient), and (iv) the verb *associated* has two interchangeable arguments (Patient A and Patient B).

Example 2 : with the sentence *Moreover, overexpression of dominant negative SHP2 blocked the protective effect of IL-6 against Dex-induced apoptosis*, the parser produces 64 parses and the output is:

```
------------------------------------------------------
Nominalization 1: overexpression
 Patient: dominant negative SHP2

Nominalization 2: effect
 Agent: IL-6
 Patient: {against} Dex-induced apoptosis

Nominalization 3: apoptosis
 Agent: Dex-induced

Verb 1: blocked (verbal sequence:  blocked ; active)
 Agent: overexpression of dominant negative SHP2
 Patient: the protective effect of IL-6 against
     Dex-induced apoptosis
------------------------------------------------------
```

In this example, the identification of all predicates and arguments are correct except for *apoptosis* where the Agent argument (*Dex-induced*) is not correct. The grammar of predicate NPs does propose an adjective as argument, but in this case *Dex-induced* is a compound adjective and was not registered as adjective in the grammar. In its present state, our system does not handle these compounds well.

Example 3: for the sentence *ET-1 expression and increased permeability may occur secondary to PKC isoform activation and may be modulated by VEGF and nitric oxide*, the parser produces 24 parses and the PredXtract output is:

```
------------------------------------------------------
Nominalization 1: expression
 Agent or Patient: ET-1

Nominalization 2: activation
 Agent or Patient: PKC isoform

Verb 1: increased  (verbal sequence:
                        increased ; passive)
 Patient: permeability

Verb 2: occur  (verbal sequence:
                       may occur ; active)
 Agent: ET-1 expression
 Agent: increased permeability
 Circumstance: {to} PKC isoform activation

Verb 3: modulated (verbal sequence:
                         may be modulated ; passive)
  Agent: VEGF
  Agent: nitric oxide
  Patient: ET-1 expression
  Patient: increased permeability
------------------------------------------------------
```

We can note: (i) the use of the modal *may* which operates on the verbs *occur* and *modulated* and which is included in the verbal sequence, (ii) the identification of the coordinate arguments of these two verbs, and (iii) an error with the Circumstance argument of the *occur* verb which is incomplete: *secondary* was ignored because the idiom *secondary to* was not recognized.

The following two short examples illustrate the presence of Circumstance roles in the verbal (Example 4-a) and nominal structures (Example 4-b) and their identification in these two structures.

Example 4-a: in this sentence *Characterization of these essential modules in transcription factors has been hampered by their low sequence homology*, the parser produces eight parses and the PredXtract output is:

```
------------------------------------------------------
Nominalization 1 : characterization
```

```
 Patient: these essential modules
 Circumstance: {in} transcription factors

Nominalization 2 : transcription

Verb 1: hampered (verbal sequence:
                  has been hampered ; passive)
 Agent: their low sequence homology
 Patient: characterization of these essential
            modules in transcription factors
-----------------------------------------------------
```

We can see that the nominalization *characterization* has two arguments: a Patient role (*these essential modules*) and a Circumstance role (*in transcription factors*) .

Example 4-b: in this other sentence *An association between cyclin D3 and the C-terminal domain of pRb2/p130 was demonstrated using the yeast two-hybrid system* the parser produces 124 parses and the PredXtract output is:

```
-----------------------------------------------------
Nominalization 1 : association
 Agent A: cyclin D3
 Agent B: the C-terminal domain of pRb2/p130

Verb 1: demonstrated (verbal sequence:
                      was demonstrated ; passive)
 Patient: an association between cyclin D3 and
            the C-terminal domain of pRb2/p130
 Circumstance: using the yeast two-hybrid system
-----------------------------------------------------
```

In this last example, we focus on Circumstance role (*using the yeast two-hybrid system*) in the verbal structure (*demonstrated*). This structure has another argument which is a Patient role (*an association between cyclin D3 and the C-terminal domain of pRb2/p130*). We can also notice the two co-agents : *cyclin D3* and *the C-terminal domain of pRb2/p130* of the nominalization *association* derived from the symetric verb *associate*.

4.2. Evaluation

PredXtract has been evaluated with a corpus of 335 Medline [1] abstracts given by biology researchers. From the 3,500 sentences of this corpus, we have selected 700 random sentences; 300 of them have been used to finalize our system and the evaluation has been done on the 400 others. In this evaluation we take into account the false positives, which are the PAS produced by the system, but which are false, and the true negatives which are the PAS that are not extracted. Because of the possibility of wrong segmentation of arguments, we have calculated two values for recall, precision and F-measure, with:
(i) [Case 1] only the true and complete arguments (the true but incomplete arguments are scored as missing arguments),
(ii) [Case 2] the true and complete arguments and the true but incomplete arguments.
The 400 sentences contain 708 nominalizations and 965 verbs; thus, nominalizations represent 42.3% of all predicates. Besides, the length of the sentences ranges from 10 to 60 words.
Table 1 presents the evaluation results for the nominalizations (N) and the verbs (V).
These results show a very small difference between values for nominalizations and verbs (at the most 0.04). So we can say that PredXtract identifies the arguments in an uniform way.
Our system obtains rather good results in the identification of arguments in case of multiple possible prepositional attachments. The main problems in parses come from long distance attachments or coordinations.
Besides, we have also calculated the recall for each sentence. We observed that there is no clear relation between sentence length (from 10 to 60 words in our evaluation) and recall values.

	N	V
True and complete arguments	508	1668
True but incomplete arguments	46	225
False arguments	86	254
Missing arguments	108	260
Case 1		
Recall	0.77	0.77
Precision	0.79	0.78
F-measure	0.78	0.77
Case 2		
Recall	0.84	0.88
Precision	0.87	0.88
F-measure	0.85	0.88

Table 1: Evaluation of verbal and nominal PAS.

4.3. Related research

Much research has been published on predicate argument structures but it is difficult to compare research because objectives are often different: as for PredXtract, it is a generic system which extracts PAS of all predicates (nominal and verbal) in the sentences processed ; the other systems, in general, aim to extract specific templates.
In biomedicine, research focuses on PAS dedicated to gene /protein interaction, where two genes or proteins are in a subject and a complement position in a proteomic relation. For example, McDonald et al. (2004) work on the specific sublanguage of gene-pathway relations, and obtain a precision rate of 89% and a recall rate of 61% with a complete parsing ; Huang et al. (2004), on protein-protein interactions, have a precision rate of 80.5% and a recall rate of 80% with a pattern-matching processing.
As Cohen et al. (2008) observe, research on nominalizations in biomedecine is very limited. Current research has rarely handled nominalizations extensively. Leroy et al. (2003) use templates built around a small set of prepositions (*of*, *in* and *by*) to capture relations with genes, proteins, gene locations, diseases, etc., they use a shallow-parsing with finite state automata and obtain 90% of precision. A specific work on PP attachments on nominalizations (Schuman and Bergler, 2006) in proteomic texts achieves good results (precision: 82%) with linguistic heuristics using information of "Specialist Lexicon" nominalizations, but the system does not produce information on the PP roles (subject, object or adjunct).

[1] Medline : a bibliographic database of biomedical information

Concerning nominalizations in other texts than biology, the first version of NOMLEX (Macleod et al., 1998) is used in information extraction (Meyers et al., 1998). The NOMBANK project (Meyers et al., 2004b) annotates automatically, semi-automatically and manually, in corpus (the Wall Street Journal Corpus of the Penn Treebank), predicate nouns (verbal, adjectival and other) with their argument relations and improves the lexical base of predicate nouns (NOMLEX-PLUS). These annotated corpora are particularly used for automatic learning.

5. Conclusion

PredXtract is a robust platform organized around the Link Parser. It parses long sentences and extracts verbal and nominal predicate argument structures. For the parsing of verbal structures as well as nominal structures, the recall, precision and F-measure values are around 0.78 without a significant difference between the three measures. This is interesting because nominalizations represent 43% of all predicates of the corpus, and thus bring a large added amount of information. These results confirm our choice to make an appropriate and effective processing of the nominalizations, as shown by Miyao et al. (2006): these authors work on similar texts and observe that their system has difficulties in processing the prepositional phrases, especially when they appear in predicate noun phrases.

As PredXtract is based on very large lexicons, it can be considered as a platform which extensively recognizes PAS, independently from the predicate type. At present, we use it for extraction of PAS in biomedical texts. To adapt it to another domain would require the addition of possible sets of specific lexical items.

To refine PredXtract outputs, we are considering extending our description of verbs with Verbnet, giving special attention to the description of diathesis alternations. In our description of verbs, the classical example with *spray* (Levin, 1993) : *Jack sprayed paint on the wall* / *Jack sprayed the wall with paint* is not taken into account at present with our system. A more precise description of these syntactic frames will also allow to improve the syntactic frames of predicate noun phrases. For a more accurate semantic description of predicates we will add semantic roles and predicate classes derived from VerbNet. In the biomedical domain, the next step will require the annotation of arguments with UMLS or other biomedical term resources.

Acknowledgments

Many thanks to Christine Brun and Bernard Jacq of LGPD-CNRS for having supplied us with their corpus of Medline abstracts tagged with gene nouns.

6. References

K. Bretonnel Cohen, Martha Palmer, and Lawrence Hunter. 2008. Nominalization and alternations in biomedical language. *PLoS ONE*, 3(9):e3158, 09.

Jing Ding, Daniel Berleant, Jun Xu, and Andy W. Fulmer. 2003. Extracting biochemical interactions from medline using a link grammar parser. In *ICTAI '03: Proceedings of the 15th IEEE International Conference on Tools with Artificial Intelligence*, page 467, Washington, DC, USA. IEEE Computer Society.

M. Gross. 1986. Lexicon-grammar: the representation of compound words. *International Conference On Computational Linguistics. Proceedings of the 11th conference on Computational linguistics*.

Jörg Hakenberg, Illés Solt, Domonkos Tikk, Luis Tari, Astrid Rheinländer, Quang Long Ngyuen, Graciela Gonzalez, and Ulf Leser. 2009. Molecular event extraction from link grammar parse trees. In *BioNLP '09: Proceedings of the Workshop on BioNLP*, pages 86–94, Morristown, NJ, USA. Association for Computational Linguistics.

M. Huang, X. Zhu, Y. Hao, D. G. Payan, K. Qu, and M. Li. 2004. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

G. Leroy, H. Chen, and J. D. Martinez. 2003. A shallow parser based on closed-class words to capture relations in biomedical text. *Journal of Biomedical Informatics*, 36:145–158.

B. Levin. 1993. English verb classes and alternation: A preliminary investigation. *The University of Chicago Press*.

C. Macleod, R. Grishman, A. Meyers, L. Barret, and R. Reeves. 1998. Nomlex: A lexicon of nominalizations. *In Proceedings of the Eighth International Congress of the European Association for Lexicography.*, pages 187–193.

Harsha V. Madhyastha, N. Balakrishnan, and K. R. Ramakrishnan. 2003. Event information extraction using link grammar. *Research Issues in Data Engineering, International Workshop on*, 0:16.

D. M. McDonald, H. Chen, H. Su, and B. B. Marshall. 2004. Extracting gene pathway relations using a hybrid grammar: the arizonarelation parser. *Bioinformatics*, 20(18):3370–3378.

A. Meyers, C. Macleod, R. Yangarber, R. Grishman, L. Barrett, and R. Reeves. 1998. Using nomlex to produce nominalization patterns for information extraction. *Proceedings of the COLING-ACL '98 Workshop on Computational Treatment of Nominals, Montreal, Canada.*

A. Meyers, R. Reeves, C. Macleod, R. Szekeley, V. Zielinska, B. Young, and R. Grishman. 2004a. The cross-breeding of dictionaries. *In proceedings of LREC-2004, Lisbon, Portugal*.

A. Meyers, R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004b. The nombank project: An interim report. *In proceeding of HLT-EACL Workshop: Frontiers in Corpus Annotation.*

Y. Miyao, O. Tomoko, M. Katsuya, T. Yoshimasa, Y. Kazuhiro, N. Takashi, and J. Tsujii. 2006. Semantic retrieval for the accurate identification of relational concepts in massive textbases. *In Proceedings of the COLING-ACL, Australia*, pages 1017–1024.

Sampo Pyysalo, Filip Ginter, Tapio Pahikkala, Jorma Boberg, Jouni Jarvinen, and Tapio Salakoski. 2006. Evaluation of two dependency parsers on biomedical

corpus targeted at protein-protein interactions. *International Journal of Medical Informatics*, 75(6):430–442, June.

Randolph Quirk, Sydney Greenbaum, Geoffrey Leech, and Jan Svartvik. 1987. *A Comprehensive grammar of the English Language.* Longman.

Jonathan Schuman and Sabine Bergler. 2006. Postnominal prepositional phrase attachment in proteomics.

D. Sleator and D. Temperley. 1991. Parsing English with a Link Grammar. *Carnegie Mellon University Computer Science technical report, CMU-CS-91-196, Carnegie Mellon University,USA.*

P. Szolovits. 2003. Adding a medical lexicon to an english parser. *In Mark Musen, editor, Proceedings of the 2003 AMIA Annual Symposium*, pages 639–643.

Lucien Tesnière. 1959. *Eléments de syntaxe structurale*. Klincksieck, Paris.

T. Wattarujeekrit, P. K. Shah, and N. Collier. 2004. Pasbio: predicate-argument structures for event extraction in molecular biology. *BMC Bioinformatics*, 5: 155.