Translating English texts into sets of predicate argument structures

Elisabeth Godbert, Jean Royaute

Translating English texts into sets of predicate argument structures

Elisabeth Godbert, Jean Royauté

Laboratoire d'Informatique Fondamentale de Marseille (LIF)
CNRS UMR 7279 - Aix-Marseille Université
Parc Scientifique et Technologique de Luminy, case 901
13288 Marseille Cedex 9
Elisabeth.Godbert@lif.univ-mrs.fr, Jean.Royaute@lif.univ-mrs.fr

1. Introduction

This paper focuses on predicate argument structures (PAS) in English texts and on the representation of events and states described in them. An event is an action either in progress or achieved. A state is a property of an object. In most cases, verbs denote actions and adjectives denote states.
Verbs and adjectives (used for example with *be*) are roots of sentences. Their nominalizations, whose syntactic patterns are more difficult to parse, allow to express the same information, except for tense marks.
We study five types of predicates: verbs, adjectives, nominalizations of verbs, nominalizations of adjectives and predicate nouns which are not related to a verb.

In most cases, verbs, adjectives and their nominalizations have the same argument relations, where arguments play precise semantic roles: they are core arguments (subjects/objects) or adjuncts. In noun phrases (NP) with nominalizations, the head noun is bound to prepositional phrases (PP) with specific prepositions which introduce arguments. For example, the NP *activation of T cells by dendritic cells* is related to the verbal form *dendritic cells activate T cells* and it is possible to insert an adjunct into the two frames, such as *in the context of pathogens*. Core arguments can also take the place of a modifier as in *T cells activation*.
Nominalizations are numerous in texts, but parsing NPs can be rather complex when they contain several nominalizations and PPs, and errors in parses are often due to incorrect prepositional attachments (Miyao et al., 2006). The first part of our work consists in a detailed study of all possible syntactic patterns for predicates, which will help us to improve prepositional attachments in parses.

The second part of our work is the translation of sentences (events and states they express) into sets of PAS expressed in an underspecified semantics. This semantics is based on three macro-roles: Agent (or Cause), Patient (or Theme) and Circumstance. The Agent is the argument which performs the action in the case of an event or to which is attached a property in the case of a state. The Patient is the argument which is involved by an action or by a state. In an active verbal form, the subject is the Agent and the object complement is the Patient, which can be introduced or not by a preposition. In passive form, roles are inverted. Circumstance is the third semantic role and corresponds to adjuncts. Thus, this underspecified semantics is at interface between syntax and semantics.

2. Typology of predicates

A typology of predicates has been defined, according to all their possible syntactic patterns. Then, predicates have been classified into seven main classes described in (Godbert and Royauté, 2010). This classification has been elaborated from scientific texts of the web, from a grammar of English and from the data of "The Specialist Lexicon" (Browne et al., 2000). Two criteria have been used to define the seven classes: (i) the role of the preposition *of* in the NP, which can mark a subject or an object complement and (ii) the role of arguments of symetric predicates, for which arguments can be exchanged. Here are a few examples from the seven classes:
- Classes 1 and 2 group together verbs accepting a direct object and the passive voice.
Heat activates electrons / Activation of electrons by heat.
John attributes human emotions to animals / Attribution of human emotions to animals by John.
- Class 6 concerns predicates with interchangeable arguments: subject and object can permute without changing the meaning.
Genes Interact with proteins / Interaction of genes with proteins / Interaction of/between genes and proteins.
Lisbon Treaty is concordant with the Czech constitution / The concordance of the Lisbon Treaty with the Czech constitution.

3. The PredXtract system

The PredXtract system is based on the Link Parser (LP) and its English native Link Grammar (LG), a variant of dependency grammars (Sleator and Temperley, 1991).
Our domain of application is biomedical text, so we have added to LG a lexicon and grammar of biological terms. The lexicon contains about 500,000 inflected forms.

In LG, links that attach verbs or nouns to any prepositional complement are generic links. In order to improve prepositional attachment and to mark the precise role of each argument of predicates, we have defined specific argument links and integrated them into the grammar.

A new grammatical module, based on argument links, has been developed for nominalizations. At the conclusion of our study of all possible syntactic patterns of nominalizations, 110 subclasses have been defined within the seven main classes mentionned in Section 2. This nominalization module contains the syntactic features of about 7,350 nominalizations, splitted into the 110 subclasses. Each nominal-

ization can accept one or more syntactic descriptions and thus can belong to several subclasses.

Besides, several modules have been developed for postprocessing the parses produced (for one sentence, often several thousands parses are produced).
The verb-adjective-noun alignment module aligns verb and adjective arguments to nominalization arguments in all parses: it integrates argument links when appropriate, identifies each verbal (or adjectival) sequence (verb with possible auxiliaries and modalities), and it identifies arguments in passive or active voice, and interchangeable arguments.
Then, for each sentence, the parses are reordered by attributing to each parse a score defined through several criteria. These criteria mainly take into account argument links in parses. For example, in the case of multiple prepositional attachments, we favor (i.e. give a higher score to) parses whose number of argument links is maximum.
At last, the syntax-semantics interface module produces for each sentence its underspecified semantic representation, close to the syntax, expressed in terms of the three macroroles Agent, Patient and Circumstance.

4. Parsing biomedical texts

An evaluation of PredXtract, for the identification of arguments of verbs and nominalizations of verbs, has been performed on a corpus of 400 random sentences from 3500 sentences of Medline abstracts. In the corpus, nominalizations represented 42.3% of all predicates. The system obtained rather good results for the identification of arguments: F-measure of approximatively 0.88 for true arguments but possibly not completely reconstituted, and 0.78 if only true and complete arguments were scored true.

Below are two examples of output of the system. Active/passive forms are noted A/P. In Ex.1 we can note (i) three nominalizations of verbs (*isolation, translation, growth*), and a nominalization of adjective (*importance*), (ii) the use of the modal *may* which operates on the verb *reflect* and is included in the verbal sequence and (iii) an error on the attachment of Circumstance with *during*, attached to *importance* instead of *involved*. According to our definition in Section 1., these PASs show one state (*importance*) and five events (*isolation, translation, growth, reflect, involved*). In Ex.2 we can note the two permutable arguments Agent A and B of *interaction*, the "That clause" of *propose* and the modal *may* on the verb *influence*.

```
=====================================================
Ex.1: Isolation of P. temperata def may reflect the
importance of specific amino acids involved in the
translation process during growth in the insect host
-----------------------------------------------------
Nominalization 1: isolation
   Patient: P. temperata def
Nominalization 2: importance
   Agent: specific amino acids involved in [..] process
   Circumstance: {during} growth in [..] host
Nominalization 3: translation
Nominalization 4: growth
   Circumstance: {in} the insect host
Verb 1: reflect (verb.seq: may reflect)(A)
   Agent: isolation of P.temperata def
   Patient: the importance of specific [..] process
Verb 2: involved (verb.seq: involved)(P)
   Patient: specific amino acids
   Patient: {in} the translation process
=====================================================
```

```
Ex.2: We propose that aberrant interaction of mutant hunting-
tin with other proteins may influence disease progression
-----------------------------------------------------
Nominalization 1: interaction
  Agent A: mutant huntingtin
  Agent B: other proteins
Nominalization 2: progression
  Patient: disease
Verb 1: propose (verb.seq: propose) (A)
  Agent: we
  That clause: that aberrant interaction [..] progression
Verb 2: influence (verb.seq: may influence) (A)
  Agent: aberrant interaction of [..] proteins
  Patient: disease progression
=====================================================
```

Adaptation to BioNLP 2011 Shared Tasks

BioNLP 2011 Tasks aimed at fine-grained information extraction (IE) in the domain of biomolecular event extraction (Kim et al., 2011). For example, from the sentence *PmrB is required for activation of PmrA in response to mild acid pH*, two events (`E1` and `E2`) must be extracted:

`E1:Pos-regulation;activation;Theme:PmrA`

`E2:Pos-reg[..];required;Cause:PmrB;Theme:E1`

Events are defined by trigger words (verbs, nouns, adjectives or complex expressions) and their arguments which are biological entities or other events. The main argument roles are "Cause" (Agent) and "Theme" (Patient).
We have participated to BioNLP Tasks. For that, we had to adapt PredXtract: complete it with specific modules for preprocessing data before parsing and for postprocessing output to extract events and write them in the right format.

5. Discussion

Much research has been done on PAS but it is difficult to compare them because objectives are often different (see for example (Johansson and Nugues, 2008) on WSJ). PredXtract includes the results of an extensive study of syntactic patterns of verbs, adjectives and nominalizations. Nominalizations are numerous in biomedical text but other research on nominalizations in biomedecine is very limited. PredXtract has been adapted in a short time to specific IE tasks for BioNLP. We now aim to use it in other domains.

6. References

A. C. Browne, A. T. Mccray, and S. Srinivasan. 2000. The specialist lexicon technical report. *Lister Hill National Center for Biomedical Communications, NLM, USA*.

E. Godbert and J. Royauté. 2010. Predxtract, a generic platform to extract in texts predicate argument structures. *Workshop "Semantic Relations", LREC 2010*.

R. Johansson and P. Nugues. 2008. Dependency-based semantic role labeling of propbank. *Proc. of the 2008 Conference on Empirical Methods in NLP*, pages 69–78.

J. Kim, S. Pyysalo, T. Ohta, R. Bossy, and J. Tsujii. 2011. Overview of BioNLP Shared Task 2011. *BioNLP 2011 Workshop Volume for Shared Task, ACL*, pages 1–6.

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. *Proc. of the COLING-ACL 2006*, pages 1017–1024.

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