Knowledge Extraction from Texts: a method for extracting predicate-argument structures from texts

Florence PUGEAULT (1, 2), Patrick SAINT-DIZIER (1), Marie-Gaëlle MONTEIL (2)

(1) IRIT-CNRS, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse FRANCE
(2) EDF, D.E.R., 1, avenue Gal de Gaulle, 92140 Clamart, FRANCE

1. Aims of the project

The general aim of our project is to improve the quality of existing systems extracting knowledge from texts by introducing refined lexical semantics data. The contribution of lexical semantics to knowledge extraction is not new and has already been demonstrated in a few systems. Our more precise aims are to:

- propose and show feasability of more radical semantic classifications which facilitate lexical descriptions by factoring out as much information as possible, enhancing re-usability of linguistic ressources. We show how the different linguistic ressources can be organized and how they interact,
- investigate different levels of granularity in the semantic descriptions and their impact on the quality of the extracted knowledge. In our system, granularity is considered at two levels: (1) linguistic: linguistic knowledge representations may be more or less precise, (2) functional: most modules of our system can work independently and thus can be used separately,
- evaluate different algorithms for extracting knowledge, taking into account efficiency aspects,
- evaluate the costs of extending our system to larger sets of texts and to different application domains.

Our project is applied to research projects descriptions (noted hereafter as RPD) where the annual work of researchers at the DER of EDF (Direction des Etudes et des Recherches, Electricité de France) is described in terms of research actions. The extracted knowledge must be sufficiently accurate to allow for the realization of the following purposes: (1) evaluation of the importance of the use of techniques, procedures and equipments, (2) automatic distribution of documents in different services, (3) interrogation, e.g. who does what and what kind of results are available, (4) identification of relations of various types between projects, (5) construction of synthesis of research activities on precise topics, and (6) creation of the 'history' of a project.

About 2.000 RPD are produced each year, each of about 200 words long. The total vocabulary is about 50.000 different words. Texts include fairly complex linguistic constructs. We also use the EDF thesaurus (encoding for nouns: taxonomies, associative relations, and synonyms, in a broad sense).

In this document, we first introduce the linguistic organization of our project, present the general form of texts and identify the type of information which must be extracted out of them. Next, we present a semantic representation for the extracted knowledge, and study in more depth the extraction of information under the form of predicate-argument and predicate-modifier structures (Jackendoff 87a, Katz and Fodor 63).

2. The overall organization of the linguistic system

Let us first introduce the way linguistic knowledge is organized. Due to space limitation, we just outline the main elements of the system. Here are the different linguistic components of our system:

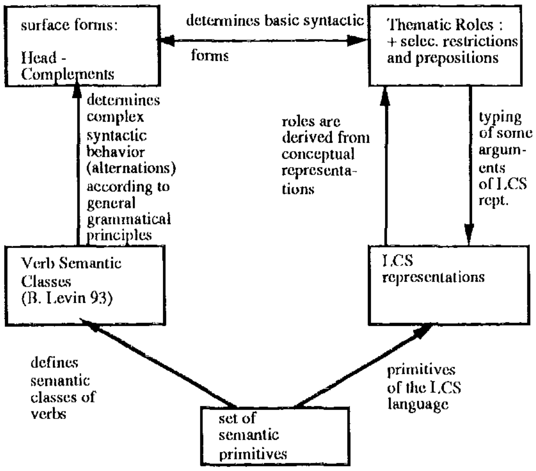
surface forms:
Head - Complements
determines basic syntactic
forms
Thematic Roles : + selec. restrictions and prepositions
determines complex syntactic behavior (alternations) according to general grammatical principles
roles are derived from conceptual representa-tions
typing of some argum-ents of LCS rept.
Verb Semantic Classes (B. Levin 93)
LCS representations
defines semantic classes of verbs
primitives of the LCS language
set of semantic primitives

Fig. 1 The General Linguistic Organization

Thematic roles (Dowty 89), (Dowty 91) paired with selectional restrictions and semantic information allow for the production or recognition of surface forms corresponding to 'basic' sentential forms. More complex forms will be treated by a system of alternations, derived from the semantic classification of verbs defined by (Levin 93).

In our approach, we consider a set of primitive elements, either general or related to our application domain, which includes notions such as being in contact with, being in spatial motion, or being the cause of. This set of primitives is designed so that it corresponds to those needed for the definition of the semantic classes of verbs, where the syntactic behavior of a verb (and thus the different ways the arguments can be distributed and should be analysed by the parser and put at the right place in the semantic representation) essentially depends on the verb's semantic nature. This approach allows for a really comprehensive treatment of predicate-argument structures because it complements the basic syntactic mappings realized from thematic roles specifications. Furthermore, this approach requires very economical lexical means since it removes a lot of idiosyncracies previously encoded in lexical entries.

We are reformulating B. Levin's work for a subset of verbs of French. Although our study is quite general, we focus primarily on verbs found in applications. Verbs of a given class have almost identical thematic distributions which are predictable from their semantics. For each of the semantic classes we have considered, we have defined a relatively small set of thematic grids, which define the 'regular' thematic distributions.

From a different perspective, we also consider that a subset of the semantic primitives we have identified are those used in the LCS, which we use in a slightly simplified way, since we do not consider for our application its deepest refinements. The efficient use of LCS for practical applications has been shown in a number of works, including (Dorr 93).

3. Semantic typology of the RPD texts

Let us first illustrate the type of text we are dealing with. Here is a standard text:

"Les mesures destructives (ou assimilables) posent toujours des problèmes concernant le faible nombre de données disponibles ou encore leur coût qui s'associe généralement à la nécessité d'une bonne précision. Il est donc nécessaire d'optimiser les campagnes de mesure pour mieux analyser les incertitudes de mesure, et, lorsque cela est possible, réduire les coûts induits. Ces problèmes sont d'autant plus difficiles à traiter que les paramètres en jeu ont des comportements non-linéaires. Il est donc nécessaire, au préalable, d'étudier les méthodes permettant de prendre en compte cette non-linéarité."

3.1 General organization of texts

A global study of these texts shows a great regularity in their overall organization. We have identified four major facets in most texts, called *articulations*. These articulations are not necessarily present altogether in a text. We have the following articulations:

- THEME, which characterizes the main purpose of the text. This articulation includes the topic of the text, and the domain on which engineers are investigating,
- MOTIVATIONS, which relate the main objectives, the needs, the goals and which explains the development of the current project.
- PROBLEMS, which correspond to the difficulties related to the current state of the art or to the limitations of certain equipments or methods.
- REALIZATIONS, which describe the different tasks required for the achievement of the project.

Articulations may cover one or more fragments of a sentence, a whole sentence or a set of sentences. They do not necessarily appear in the order they have been defined here. The decomposition of texts in articulations defines the **pragmatic level**. We view the articulations as defining semantic fields. The above text can be decomposed as follows:

[**theme** [les mesures destructives]],
[**motivations** [optimiser les campagnes de mesure pour mieux analyser les incertitudes de mesure, et, lorsque cela est possible, reduire les coûts induits.]],
[**problems** [[posent toujours des problemes concernant le faible nombre de données disponibles ou encore leur coût qui s'associe generalement à la necessité d'une bonne precision], [problèmes sont d'autant plus difficiles à traiter que les paramètres en jeu ont des comportements non-linéaires.]],
[**realizations** [étudier les méthodes permettant de prendre en compte cette non-linéarité.]]].

For this level, we have implemented a method which permits the identification of the different articulations of a text. This problem is divided into two sub-problems: (1) identification of the articulations, and (2) extraction of relevant sentence fragments from the original text.

A study of the RPD texts has shown that these four articulations can relatively easily be identified by means of specific terms or constructions. Let us call these terms or constructions *articulation triggers*. Articulation triggers belong to different linguistic domains:

(1) *lexical*, where triggers are just words, e.g. 'devoted to', 'in the context of', 'propose', for THEME,

(2) *grammatical*, where triggers can be phrases, or related to grammatical information (such as tense and aspect, e.g. 'in the past years', 'since 1989', for THEME), or verbs or nouns of certain semantic class, e.g. verbs of volition, of creation (Levin 93),

(3) *discursive*, where triggers are mainly propositional connectors such as 'therefore', 'because', etc.,

(4) *pragmatic*, where the relative positions of sentences and more generally, the physical form of texts (e.g. enumerations) can determine articulations.

The next stage is to extract those portions of text which are relevant for the articulation considered. Since the linguistic treatements of this first level are necessarily superficial, we must carefully discard irrelevant portions of texts. This approach has been modelled by means of *extraction rules*, which specify words and constructions to skip and which delimit zones of texts to be extracted. Evaluation of results is given in fig. 2 in the annex.

3.2 Identification of knowledge to be extracted

Let us now concentrate on the nature of the semantic information which should be extracted by the system. We have identified three types of information:

- *general nominal terms* (e.g. 'methods', 'data'), and *specific nominal terms* belonging to technical domains,
- *states or actions* in which these terms are involved,
- *general roles* played by these terms in actions or states.

Roughly speaking, the first class identifies arguments, the second class defines predicates, while the third one introduces the notion of semantic roles such as *thematic roles*. This latter level is of a crucial importance in knowledge extraction because it avoids making incorrect interpretations on the role of an argument with respect to the action or state being described. This level is called the **linguistic level**.

The level of granularity we are considering in this project suggests us to group predicates with a close meaning into a class and to represent them by the same predicate name, viewed as a primitive term. For example, we have terms which express the notion of *definition* (e.g. define, specify, describe, identify, qualify, represent) or the notion of *building* (e.g. assemble, build, compile, develop, forge) as defined in B. Levin's work. However, for a relatively small number of classes, in particular for those classes of predicates which denote complex actions and for those which exhibit a high degree of incorporation (Baker 88), where incorporated knowledge needs to be made more explicit, it may be necessary to use a more conceptual type of representation. We want to investigate the use the Lexical Conceptual Structures (LCS) (Jackendoff 87, 90) which match very well with the planned uses of the extracted knowledge on the one hand, and with the notion of thematic roles on the other hand. Let us call it the **conceptual level**. This paper being mainly devoted to the linguistic level, this level will not be investigated here.

4. The linguistic level

4.1 Identification of predicative terms

Predicative terms characterize states or actions. The goal at this stage is to be able to determine in a way which is as systematic as possible which terms are predicative in the RPD texts. A priori, verbs denoting states or actions and prepositions are considered to be predicative terms. Nouns are slightly more difficult to treat. The EDF dictionary includes the specification of nouns derived from verbs. We consider that these nouns are predicative. A few nouns, not derived from verbs are also predicative, such as algorithm, sort or departure, these are identified so far by hand. They may be later semantically classified as describing, for example, actions or events.

4.2 Identification of relevant predicates and arguments in texts

The second aspect of the linguistic level is the identification of predicates and related arguments which are sufficiently relevant to be extracted. Relevance can be defined a priori and once for all or may depend on the text. The relevance of a term can be defined according to several criteria:

(1) *genericity*, terms defining a research action, a realization, or a problem such as: define, improve, implement, test, evaluate and explore are of much interest. At this level, it is most useful to use B. Levin's verb classification to determine relevance.

(2) *specialization*, corresponding to very precise terms describing a material, an equipment, a method or a system. Specialized terms can be defined a priori from the thesaurus by extracting the most specialized terms.

(3) *local importance*, where importance in a text is explicitly marked, for example, by a construction such as 'it is important to...' or by a negation.

4.3 Representing predicate arguments and modifiers by means of thematic roles

The relationship between a predicate and one of its arguments can be represented by a thematic role. Thematic roles do confer a much stronger meaning to predicate structures, in particular when thematic roles have a relatively precise meaning. Thematic roles can be defined in a more refined way than the usual definitions. From that perspective, our claim is that thematic roles can form the basis of a good and stable general descriptive semantics of predicate-argument relationships. Thematic roles have then a *conceptual dimension*, and not only a linguistic one. However, they must not be confused with the conceptual labels of the LCS. Thematic roles must remain general; they form a bridge between conceptual representations and syntax. Fig. 3 shows the thematic roles we consider.

We consider here an extended use of thematic roles since they are also assigned to predicate modifiers, realized as prepositional phrases or as propositions, in order to represent in a more explicit and uniform way essential arguments and modifiers, since they all play an important role in the semantics of a proposition.

The general form of a semantic representation introduces two functions for thematic roles:

(1) *an argument typing function*:

predicate_name(..., role$_i$: {arg$_i$ }, ...)

(2) *a predicate modifier typing function*, where a predicate is marked by a thematic role, if the modifier is a predicate:

role$_j$: predicate_name(..., role$_k$: {arg$_k$ }, ...)

The arg$_i$ are fragments of texts (NPs and PPs), which may be further analyzed in a similar way, if necessary. For example, a sentence such as:

John got injured by changing a wheel

is represented by:

injured(theme : {john}) ∧ causal theme :
change(agent: {john} , theme : {wheel}).

If in an articulation, we only extract an NP, it is represented as an argument as follows:
arg({ fragments of text corresponding to the NP }).
and no thematic role is assigned to it. The general representation of an articulation is then:

[**articulation_name**,
[extracted text from pragmatic level],
partial predicate-arg representation]

The result of the parse of our sample text is given below.

[[**theme** [les mesures destructives (ou assimilables)]
arg: {mesures destructives}] ,
[**motivations** [optimiser les campagnes de mesure pour mieux connaitre, voire ameliorer, les incertitudes de mesure, et, lorsque cela est possible, reduire les coûts induits.]
optimise(_ , Incremental beneficiary theme:
{campagnes de mesure}) ∧
goal: (analyze(_ , holistic theme:
{incertitudes de mesure}) ∧
reduce(_ ,incremental victim theme:{coûts}))] ,
[**prob!ems** [[posent toujours des problemes concernant le faible nombre de donnees disponibles ou encore leur coût qui s'associe generalement à la necessité d'une bonne precision.] [problèmes sont d'autant plus difficiles à traiter que les parametres en

jeu ont des comportements non-lineaires.]
arg: ({faible nombre de donnéas}, {coût},
{comportements non-linéaires})] ,
[**realizations** [étudier les méthodes permettant de prendre en compte cette non-linéarite.]
study(_, general theme: {methods})]].

4.4 Parsing and assigning thematic roles

Let us now show how our parser works and how thematic roles are concretely assigned to arguments. For that purpose, we introduce three main criteria:

(1) the semantic class of the predicative term where thematic grids are given,

(2) the semantic type of the preposition, if any, which introduces the argument, we also have defined thematic grids for prepositions,

(3) the general semantic type of the head noun of the argument NP. Semantic types are mainly defined from the semantic fields given in the EDF thesaurus.

These criteria are summarized in fig. 4 at the end of this document. These criteria are implemented by means of thematic role assignment rules.

The parsing of the RPD texts works independently on each fragment of text associated with each articulation (referencial aspects will be considered later). We have the three following stages:

(1) Identification of predicates and arguments: due to the complexity of texts, a partial analysis is the only possible and efficient solution. We have a grammar that identifies basic verbal constructions, nominal constructions. The parser works bottom-up and identifies maximal structures which are not ambiguous.

(2) Thematic role assignement: The assignment procedure considers each thematic role in a thematic grid and searches for a nominal or propositional structure to which the thematic role can be assigned. This assignment is based on the thematic role assignement rules. The general form of a thematic role assignment rule is the following:

```
assign_role(<name of role>,
        <grammatical form of predicate>,
        <grammatical form of argument>) :- <unification or
subsumption constraints on semantic features>.
```

This is illustrated as follows, where grammatical forms (xp) are given in Login form (Aït-Kaçi and Nasr 86), following the TFS approach:

```
assign_role(effective_agent,
        xp(syntax => syn(cat => v), semantics =>
         sem( pred => yes, relevance => yes)),
        xp(syntax => syn(cat => n), semantics =>
                sem( pred => no,
        sem_type => tsem( semp => X )))) :-
                subsumed(X, [human, technical]).
```

This process can be applied recursively on those arguments which contain predicates. The depth of recursion is a parameter of the system.

(3) Semantic representation construction. At this level, deeper representations (such as the LCS) can be used.

Conclusion

The novelty of our approach with respect to knowledge extraction can be summarized as follows:

(1) We have defined *three levels of knowledge representation* (pragmatic, linguistic and conceptual), which are *homogeneous*, expressed within a *single, incremental formalism*, incremental in the sense that knowledge extracted at an outer level is refined at a deeper one, and that representations support *partial information.*

(2) We have defined simple *methods for extracting relevant terms* in texts, using a thesaurus.

(3) We show that the syntactic alternations given in Levin's work complement the basic syntactic forms generated from thematic roles. These semantic classes of verbs, because of their semantic basis and because of the way they are defined are *a very powerful tool for assigning correctly thematic roles* to predicate argument in a large number of syntactic forms.

(4) The different types of data and the level of granularity at which they are considered establishes linguistic levels of descriptions which correspond to a certain descriptive reality and to a certain autonomous and homogeneous level of semantic representation.

Acknowledgements

We are very grateful to Marie-Luce Herviou, Palmira Marrafa and Sophie Daubèze for discussions on this project. We also thank Martha Palmer and Bonnie Dorr for several discussions and for introducing us to B. Levin's work. This project is funded by EDF-DER.

References

Aït-Kaçi, H., Nasr, R., LOGIN: A Logic Programming Language with Built-in Inheritance, *journal of Logic Programming*, vol. 3, pp 185-215, 1986.

Baker, M. C., *Incorporation, A Theory of Grammatical Function Changing*, Chicago University Press, 1988.

Blosseville MJ, Hebrail G, Monteil MG, Penot N, Automatic Document Classification: Natural Language Processing, Statistical Analysis and Expert System Used Together, *ACM SIGIR*, Copenhaguen, June 1992.

Dorr, B., *Machine Translation: a View from the Lexicon*, MIT Press, 1993.

Dowty, D., On the Semantic Content of the Notion of Thematic Role, in *Properties, Types and Meaning*, G. Cherchia, B. Partee, R. Turner (Edts), Kluwer Academic Press, 1989.

Dowty, D., Thematic Proto-roles and Argument Selection, *Language*, vol. 67-3, 1991.

Grimshaw, J., *Argument Structure*, MIT Press, 1990.

Jackendoff, R., The Status of Thematic Relations in Linguistic Theory, *Linguistic Inquiry* 18, 369-411, 1987.

Jackendoff, R., *Conciousness and the Computational Mind*, MIT Press, 1987.

Jackendoff, R., *Semantic Structures*, MIT Press, 1990.

Katz, J. J., Fodor, J. A., The Structure of a Semantic Theory, in *Language* 39, pp. 170-210, 1963.

Levin, B., *English Verb Classes and Alternations*, the University of Chicago Press, 1993.

articulations	fully correct extraction	partly correct extraction	incorrect extraction
<THEME>	86%	11,5%	2,5%
<MOTIVATIONS>	70%	21,5%	8,4%
<PROBLEMS>	61%	33,5%	5,5%
<REALISATIONS>	46,5%	30%	23,5%

Fig. 2 evaluation of level 1

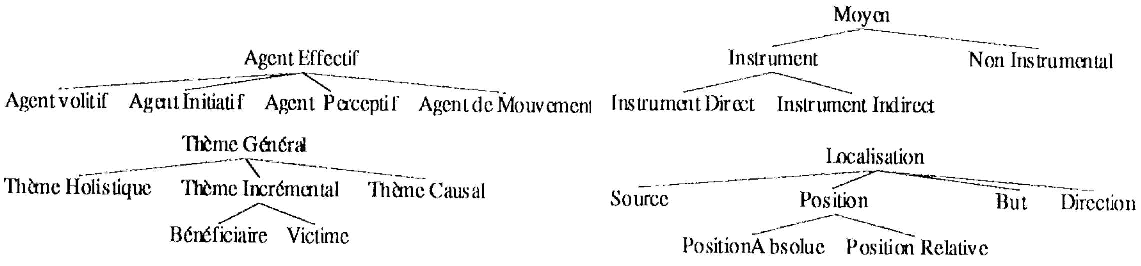

Fig. 3 The thematic role hierarchy (in French)

Thematic role	**Semantic class of predicate**	**Selectional restr. on argument + prep**	**Examples**
Effective Agent (ae)	characterize creation and transfo. continue service transfer of possession searching, etc.	human	définir, représenter, créer, réaliser, continuer, poursuivre, aider, collaborer, donner, échanger, rechercher, résoudre, etc.
Volitive agent	volition obligation	human	vouloir, désirer, devoir, obliger, nécessiter.
Initiative agent	allowing decision	human | technical	favoriser, permettre, conduire, décider, diriger, mener.
Perceptive agent	knowledge	human	savoir, connaitre.
Agent of Movement	continue	concrete_element | human	étendre, poursuivre,
Theme	searching obligation tranfer of possession attaching, etc.	- animate | technicale | human	explorer, observer, devoir, obliger, donner, échanger, attacher, chaîner, etc.
Means	creation and transfo. characterize, etc.	prep: avec, en, par	construire, réaliser utiliser, spécifier (par).
Localization	moving attaching	place (spatial loc.) temporal (temporal loc.) abstract | technical (abstract loc.) prep: dans, sur, de, etc.	aller, venir, attacher, chaîner, relier (à).
Identifier	identification	proper_noun | profession.	baptiser, nommer.
Accompaniement	service attaching	animate prep: avec	collaborer, participer, attacher (avec), unir.

Fig . 4 Sample of the organization of thematic roles w.r.p. to semantic classes of verbs, selectional restrictions and prepositions.