Elie Naulleau, Electricité de France, DER et ELI
Marie-Gaëlle Monteil, Electricité de France, DER
Benoît Habert, ELI

Process Terms

Abstract

In this paper, we describe a method to build terminological selection filters using an existing terminological resource, the EDF thesaurus. We first proceed to a linguistic enrichment of the available terminology, using NLP software and an existing dictionary. This process leads us to manipulate terms as decorated syntactic trees. Then we try to model what a term is in the EDF thesaurus only by considering the relationship between nouns and adjectives, using a relational analysis clusterization software from IBM. The result is the induction of rich patterns, based on morphological, syntactic, semantic and domain information. These patterns help to filter and select term candidates from a text, and so, make easier the updating of our thesaurus.

1. Introduction

EDF, the French electricity company, uses its terminologies in a large set of applications such as automatic indexing, information retrieval, technological watch and automatic dissemination of information. To follow the rapid terminological evolution in technological domains, EDF has developed, in the last 25 years, an activity of creation, management and updating of its terminologies. Today, this activity is supported by internal researches,[1] by an in-house documentary expertise and by the results of Europeans projects such as GRAAL (Grammars which are Reusable for Automatic Analysis of Language)[2] and TRANSTERM (integration of terminologies in NLP software) in which EDF is involved.

Our work consists in elaborating a terminological model taking into account syntactic and semantic realities. Such modelization of the term is useful for the acquisition, the management and the use of terminologies. For example, it permits to define filtering criteria for terminology acquisition from corpora. It also makes the terminological resources more easily usable in NLP applications (terminology is thus considered as a part of the lexicon). This is crucial as, up to now, terms were currently recorded as flat strings characters in the EDF thesaurus and the nominal phrases appearing in documents often do not match them,

because the control of lexical and syntactic variation that occurs in texts needs a linguistic structured representation of terms.

We present in this paper an experiment we have achieved in order to select term candidates in texts, by recycling and taking advantage of an existing terminological resource, the EDF thesaurus. We describe how we have refined the thesaurus with linguistic information (section 2), then, we explain how the enriched terminological entries can help us to build selection filters (section 3) and so, help us to update the thesaurus.

1.1 Underlying hypothesis

These three underlying hypothesis have guided our experiment:

a) From the point of view of knowledge acquisition, textual data is a source of linguistic description: textual processing based on syntax can reveal semantic pattern and classes (as shown by [Sager & al -87], [Habert & Fabre, -95] and [Habert & al, -95]), and textual processing based on the combination of syntax and semantic may reveal conceptual information [Habert & Nazarenko, -96], [Habert & al,96].

b) A terminology is a knowledge-based object, which is not naturally and immediatly visible through the language. It comes from an effort of normalization. For this reason the search of terminological information with a computer cannot come from a conventional reading of the text, but may arise from specific investigation and organisation processes, such as syntactic normalization of phrases, normalization of variants.

c) For a large part of terms and term candidates,[3] it is possible to distinguish them from common nominal phrases, if they are described as rich linguistic objects especially using semantic information.

1.2 An approach derived from an harrissian framework

These hypothesis, especially the first one, has lead us to choose a harrissian framework, which applies to the description of a sublanguage associated to a knowlegde or activity domain. The characterization of such a sublanguage requires the normalization of its sentences, bringing out the main operators and their argument classes,[4] which are

semantically homogenous (see [Dachelet, -94]). This approach enables first an inductive updating of properties from a corpus; in a second step, it enables to model the linguistic objects with the properties which have been inducted. Our protocol of experience is however quite different from the harrissian framework: we have limited the scope of our experience to nominal phrases, because we use the EDF thesaurus, which provides the terms without any context. In addition the nominal phrases we are working on are provided by an automatic analysis process and have undergone an automatic linguistic enrichment. The thesaurus is constitued of about 20,000 certified terms, which could reveal us their linguistic properties. Moreover, it is organized in knowledge domains, which are divided in semantic fields, defining sublanguage areas.

2. The enriching process

For the present experience, the whole terms of the EDF thesaurus have constituted the input text given to the parser AlethGram (AlethGram is a robust grammar coming from the GRAAL project; it runs with a pattern matcher called AlethMPM developed by Gsi-Erli), e.g. a list of certified terms. The parser makes a morphological analysis (categorization) and a syntactic analysis providing in output syntactic trees, representing the parsed terms. We then have enriched these trees with various linguistic information, such as :

- morphological information (suffixes of adjectives (-IQUE for "linguist-ique"), and suffixes of nouns (-AGE for "surmen-age"), derivated forms in other lexical categories (interroger (verb), interrogeable (adjective), interrogation (noun)),
- syntactic information (predicativity of the nouns (does a noun accept an argument), syntactic constructions for nouns (how many and what kind of argument does a noun accept), struc-ture of the term),
- semantic information (all possible referent types of a noun are associated to it (is a noun an abstract noun, a substance noun, a artefact noun, ...? We are using about 80 referent types defined in a hierarchy), semantic tags are associated to the adjectives. These tags are defined for a documentary relevance, we will not use them here).

The morphological suffixes and affixes are imported from [Guilbert,

-70]. The other pieces of information are extracted directly (lexical entries, derivated forms, syntactic constructions, predicative nouns) or indirectly (referent type of nouns) from EDF internal projects, or have been inspired from existing projects (as Wordnet [Miller, -93]), or have been constituted manually (semantic values for adjectives).

So, we can manipulate nominal phrases, not as flat forms (character strings), but as syntactic trees,[5] adorned with various linguistic data. The term "brûleur à pulvérisation mécanique" (mecanical pulverization burner) is represented as shown in figure 1, at the end of the process.

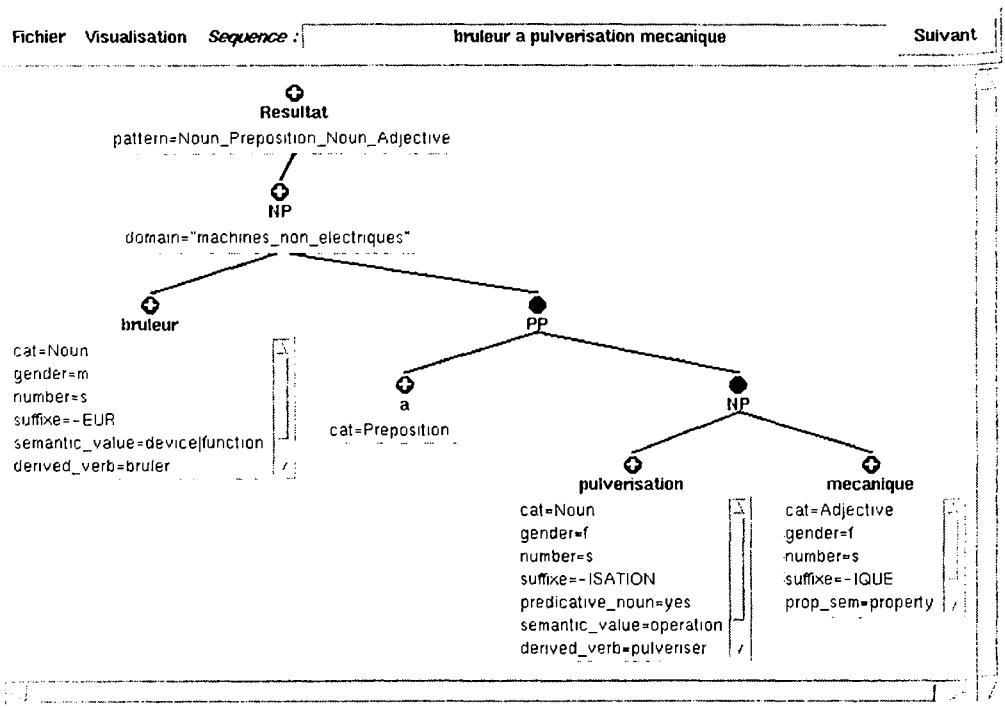

Figure 1: "an enriched parse tree for the term "brûleur à pulvérisation mécanique"

3. The semantic and syntactic behavior of the adjectives in the EDF thesaurus

Terms can be described as adjunctions and combinations of elementary modifiers over a nominal head. Among the two main kinds of elementary modifiers : "~ adjective" and "~ preposition noun", we have chosen the adjective one to characterize the term. It turned out that this relation between a noun and its adjective is an interesting starting point to build

selection filters for term candidates; because the relatively simple noun-adjective relation allows to point out basic referential alterations and semantic properties. These peculiarities, which have been used by [Assadi & Bourigault, -95] to model knowledge from texts, define relevant and notable selection restrictions for the terminological and documentary usage.[6]

3.1 Clusterization with an relational analysis software.

To bring out what a term is from the point of view of the noun-adjective relation, we have extracted all the [noun, adjective] couples from the enriched thesaurus. We have then applied the clustering software Tewat.[7] The use of a clustering method has facilitated the intepretation of the various pieces of information attached to each term, distributed in as many dimensions, and finally synthesized in the form of classes. The individuals were the adjectives; the modalities used to define the classes were the suffixes of the adjectives and the nouns, the referent types of the nouns, the sub-domain code associated to the term in the thesaurus and the syntactic pattern of the term. The result is a set of classes of adjectives defined for certain modalities.

3.2 The building of selection filters

Three classes computed by Tewat are shown below (see next page):

The result 1 is a class made of all the possible adjectives for the unique name *acide* (acid). The thesaurus sub-domains modalities coming with the class are 'composé organique' (organic compound) and 'cytologie' (cytology). The referent type of the noun is: SUBSTANCE; the suffix of the adjective is: -IQUE. The syntactic pattern modality concerned is: Noun_Adjective.

The result 2 is a class made of the whole possible adjectives for the unique name *roche* (rock). The thesaurus sub-domain coming with this class is 'roche' (rock), the referent type of the sole noun is: OBJECT or SUBSTANCE; multiples suffixes for the adjective are possible : -IQUE, -U, -IBLE, -EUX, ... The syntactic pattern modality concerned is : Noun_Adjective.

The result 3 is a class of adjectives functioning with certain nouns. The nouns coming with the class are *élevage, production*. The thesaurus sub-domains modality belong to the 'agriculture' domain. Referent type of

the nouns are ACTIVITY and OPERATION; nouns suffixes are : -TION and -AGE, and adjective ones are : -IN, -OLE and -IF.

	Result 1	Result 2	Result 3
Adjective class	{palmitique, lactique, benzoïque, oxalique, stéarique, phtalique, butyrique, acétique, formique, ascorbique, carboxylique, tartrique, picrique, oléique, sulfonique, citrique, ribonucléique, désoxyribonucléique, urique, gras, aminé}	{microlitique, détritique, plutonique, magmatique, tendre, métamorphique, sédimentaire, igné, siliceux, cristallophylle, microgrenu, dur, combustible}	{ovin, porcin, bovin, avicole, intensif, extensif, animal}
Suffixe modalities for adjectives	-IQUE	-IQUE, -U, -IBLE, -EUX,...	-IN, -OLE , -IF
Suffixe modalities for nouns	-	-	-TION , -AGE
Noun modalities	acide (acid)	roche (rock)	élevage (stock-farming), production
Referent type of the noun	SUBSTANCE	SUBSTANCE/ OBJECT	ACTIVITY/ OPERATION
Syntactic pattern	Noun adjective	Noun adjective	Noun adjective
Sub-domain modality	'composé organique' (organic compound), 'cytologie' (cytology)	'roche' (rock)	'agriculture'

When modalities are turned into constraints, such classes are inter-preted as filtering patterns with fuzzy limits (the classes are built from the likeness and not from the identity of the modalities). The filter acts either as a list of conditions to be satisfied, or as a simple equation: the unknown parameter can be deduced while the other ones are known. Thus, there are several ways of interpreting a sequence with the resulting filter, because it can match it in several ways:

a) It is possible to predict the domain or the sub-domain of the nominal phrase when its description corresponds to the constraints given by the filter including lexical constraints (presence of a given word or an affixe). Example: *acide chlorhydrique* (hydrochloric acid) has been found in a text; its syntactic flat pattern is Noun_Adjective, the adjective ends with "-IQUE", so its possible thesaurus sub-domains are: 'composé organique' (organic compound) or 'cytologie' (cytology).

b) The same situation occurs, without lexical adequation but with referent type adequation of the noun and some morphological adequation of the adjective; here, it is not possible to predict the sub-domain; however the sequence may be considered as a potential term. Example: *éther sulfurique* (sulphuric ether), *gaz sulphuric* (sulphuric gaz) – gaz and ether are kind of SUBSTANCE and *sulphurique* ends with "-IQUE".

c) If the domain of the text is known, the nominal phrase may or may not be a term candidate. For example: *huile synthétique* (synthetic oil) has been found in a paragraph dealing with the 'matériaux' (materials) thesaurus domain. It could be a potential term because it has matched an already validated filter (a SUBSTANCE followed by an "-IQUE" adjective), used in other close domains (as it concerns materials). Here is another example: *sélection ovine* (ovine selection) has been found. It matches the filter because *sélection* is an OPERATION and the adjective belongs to the list mentioned by the class. It could be a candidate, above all if the text deals with agriculture.

4. Conclusion

We have used an existing thesaurus to bring out rich patterns made of morphological, syntactic, semantic and domain pieces of information. These patterns can act as selection filters and generally depend on a domain of activity. The filters consist in symbolic rules (lists of linguistic conditions to be statisfied) that have been obtained by a statistical process: a clustering method. There has been induction of information from a previously enriched linguistic material, the terms of the thesaurus.

5. Future work

When several referent types are possible for a noun, they are all associated to it. For example the noun *base* (base) can be a SUBSTANCE (chemistery), an ARTEFACT (database), a BUILDING (military base) or an ABSTRACT noun (geometrical location, arithmetic or linguistic radix, ...), and so on. This damages the quality of the classes and their intelligibility: the classes built with ambiguous nouns are not easily interpretable as filters, even if the thesaurus sub-domain gives a clue. This is the reason why the sole results we show are based on non ambiguous nouns. For the time being, we are working on a disambiguation system, which is able to assign semantic tags for nouns and adjectives (going beyond the values of morphological suffixes) by considering their context. The system is being trained on the thesaurus and a reference corpus. It is evident that the disambiguation rules will depend on them. However, in our lexical database, the lexemes are recorded with the conditions of choice of their meaning. So we are gathering enough contexts to ensure a certain genericity of the rules for the most frequent ambiguous nouns.

The disambiguation process applied before the clustering method will enhance the interpretability of the classes, and so, the accuracy of the resulting filters.

We will then test the efficiency of the filters on a corpus and compare the extracted nominal phrases with nominal phrases coming from a simple extraction. If the results look good, we will extend the scope of the filters taking into account the "~ preposition noun" elementary modifier.

Notes

1. See [Pugeault,95], [Sta,95].
2. The French participants for GRAAL are EDF, Aérospatiale, Gsi-Erli (coordinator), Renault. The French participants for the TRANSTERM project are Gsi-Erli (coordinator), Aérospatiale, EDF.
3. We define a term candidate as a noun phrase which has been extracted from a text, and which could be considered as a potential term because it seems to denote an object of the domain described in the text.
4. For instance, sets of compatible arguments for some given predicative forms.
5. The tree representation expresses the syntactic relations between its compounds in an unambiguous way, insofar as the parser gives a correct analysis of the phrase.

6. For instance, *Tarif bleu* (blue rate) can be found in the EDF thesaurus. *Tarif* admits a degree modification (hight, low, ...), but generally does not admit a color modification. In the pattern Noun(abstract) Adjective(color), the adjective has only a distinctive value and could produce a metaphorical effect. From a terminological point of view, this phenomenon has to be noticed: it could establish a generic relation (hyperonym) bewteen *tarif bleu* and *tarif*.
7. Clustering method based on the relationnal data analysis [Michaud, -87], [Condorcet, -85], implemented in the Tewat software by IBM.

References

Assadi H., Bourigault D., (1995), "Classification d'adjectifs extraits d'un corpus pour l'aide à la modélisation de connaissances", in *Proceedings JADT'95*, Rome.

Church, K., Hanks, P., (1990), "Word Association Norms, Mutual Information, and Lexicography", *Computational Linguistics* 16(1):22–29, march 90.

Condorcet (de), A. (1785), *Essai sur l'application de l'analyse de la probabilité des decisions rendues à la pluralité des voix*, Paris.

Dachelet, D., (1994), *Sur la notion de sous-langage*, Thèse de doctorat en sciences du langage, Université Paris VIII.

Daille, B. (1994), *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*, Thèse de doctorat en informatique, Université Paris 7.

Gouadec, D. (1992), *Terminologie et Terminotique - Outils, modèles et méthodes*, Ed. La Maison du Dictionnaire, Paris.

Guilbert, L. (1970), "Fondements lexicologiques du dictionnaire - de la formation des unités lexicales", in *Grand Larousse de la Langue Française*.

Habert, B., Fabre, C. (1995), Simplifying nominal parse trees to find semantic types in corpus, in *Proceedings ALLC-ACH*.

Habert, B., Barbaud, P., Dupuis, F., Jacquemin, C., (1995), "Simplifier des arbres d'analyse pour dégager les comportements syntactico-sémantiques des formes d'un corpus" in *Cahiers de grammaire* N°20, Université Toulouse le Mirail.

Habert, B., Nazarenko, A., (1996), "La syntaxe comme marche-pied de l'acquisition des connaissances : bilan critique d'une expérience", in *Proceedings JAC'96*

Habert, B., Naulleau E., Nazarenko A., (1996), "Symbolic word classification for medium-size corpora", accepted for COLING'96, Copenhagen.

Michaud, P. (1987), "Condorcet, a man of the avant-garde", *Journal of Applied Stochastic Model and Data analysis*, Vol 3, No 2.

Miller, G.A., (1993), "Introduction to Wordnet : an on-line lexical database", "Nouns in Wordnet : a lexical inheritance system", in 5papers.ps (ftp://clarity.princeton.edu)

Pugeault F, Saint-Dizier P., Monteil MG, "Knowledge Extraction from texts : a method for extracting predicate argument structures from texts", in *Proceedings COLING'94*, Kyoto

Sager, N., Friedman C., Lyman, MS., (1987), "Medical Language Processing : Computer Management of Narrative Data", Addison-Wesley.

Sta, J.D., (1995), "Document expansion applied to classification : weighting of additional terms", in *Proceedings ACM-SIGIR'95*, Seattle.