A step towards the detection of semantic variants of terms in technical documents

Thierry Hamon and **Adeline Nazarenko**
Laboratoire d'Informatique de Paris-Nord
Université Paris-Nord
Avenue J-B Clément
93430 Villetaneuse, FRANCE
thierry.hamon@lipn.univ-paris13.fr
adeline.nazarenko@lipn.univ-paris13.fr

Cécile Gros
EDF-DER-IMA-TIEM-SOAD
1 Avenue du Général de Gaulle
92141 Clamart CEDEX, FRANCE
cecile.gros@der.edfgdf.fr

Abstract

This paper reports the results of a preliminary experiment on the detection of semantic variants of terms in a French technical document. The general goal of our work is to help the structuration of terminologies. Two kinds of semantic variants can be found in traditional terminologies : strict synonymy links and fuzzier relations like *see-also*. We have designed three rules which exploit general dictionary information to infer synonymy relations between complex candidate terms. The results have been examined by a human terminologist. The expert has judged that half of the overall pairs of terms are relevant for the semantic variation. He validated an important part of the detected links as synonymy. Moreover, it appeared that numerous errors are due to few mis-interpreted links: they could be eliminated by few exception rules.

1 Introduction

1.1 Structuring a terminology

The work presented here is a part of an industrial project of Technical Document Consultation System (Gros et al., 1996) at the French electricity company EDF. The goal is to develop tools to help a terminologist in the construction of a structured terminology (cf. figure 1) providing :

- **terms** of a domain, i.e. simple or complex lexical units pointing out accurate concepts in a technical document, (Bourigault, 1992);
- semantic links such as the see-also relation.

This can be viewed as a two-step process. The **candidate terms** (i.e. lexical units which can be terms if a domain expert validates them) are first automatically extracted from the technical document with a Terminology Extraction Software (LEXTER) (Bourigault, 1992). The list of candidate terms is then structured into a semantic network. We focus on the latter point by detecting semantic variants, especially synonyms.

ligne aérienne (overhead line)
  See_also : Départ aérien (overhead outlet)
  Synonym : Liaison électrique aérienne (overhead electric link)
Ligne simple (single circuit line)
  Is_a : Ligne aérienne (overhead line)
Ligne multiterne (multiple circuit line)
  Is_a : Ligne aérienne (overhead line)
  Synonym : Ligne double (double circuit line)

Figure 1: Example of a structured terminology in the electric domain.

In order to build a structured terminology, we thus attempt to link candidate terms extracted from a French technical document[1]. For instance, from synonyms such as *matériel* (equipment) / *équipement* (fittings), *marche* (running) / *fonctionnement* (working) and *normal* (normal) / *bon* (right), we infer a synonymy link between candidate terms *matériel électrique* (electric equipment) / *équipement électrique* (electrical fittings) and *marche normale* (normal running) / *bon fonctionnement* (right working).

[1] As the terms used in this paper have been extracted from French documents, their translation, especially for the synonymy, does not always show the same nuance than originally.

modèle (model) : $<1>$ canon (canon), étalon (standard), exemplaire (copy), exemple (example), plan (plan)
$<2>$ sujet (subject), maquette (maquette)
$<3>$ héros (hero), type (type)
$<4>$ échantillon (sample), spécimen (sample)
$<5>$ standard (standard), type (type), prototype (prototype)
$<6>$ maquette (model)
$<7>$ gabarit (size), moule (mould), patron (pattern)

Figure 2: Example of a word entry from the dictionary ***Le Robert***.

1.2 Using a general language dictionary for specialized corpora

As domain specific semantic information is seldom available, our aim is to evaluate the relevance and usefulness of general semantic resources for the detection of synonymy between candidate terms.

For this study, we used a French general dictionary *Le Robert* supplied by the Institut National de la Langue Française (INaLF). It provides synonyms and analogical words distributed among the different senses (cf. figure 2) of each word entry. It is exploited as a machine-readable synonym dictionary.

We use a 200 000 word corpus about electric power plant. Its size is typical of the technical documents. It is very technical if one considers the dictionary lemma coverage for this corpus (45%). Concerning two other available documents dealing with software engineering and electric network planning, the dictionary lemma coverage is respectively of 65% and 57%. In that respect the chosen corpus is the worse case for this experiment.

The present corpus has been analyzed by the Terminology Extraction Software LEXTER which extracted 12 043 candidate terms (2 831 nouns, 597 adjectives and 8 615 noun phrases). Each complex candidate term (*ligne d'alimentation*, supply line) is analyzed into a head (*ligne*, line) and an expansion (*alimentation*, supply). It is part of a syntactic network (cf. figure 3).

2 Method for the detection of synonymous terms

The terminological variation include morphological (flectional, derivational) variants, syntactic variant (coordinated and compound terms) but also semantic variant (synonyms, hyperonyms) of controlled terms. In this experiment, we attempt to infer synonymy links between candidate terms.

2.1 Semantic variation and synonymy relation

Semantic variation The semantic variation includes relations (e.g. synonymy and see-also) between words of the same grammatical category, even if one may also take into consideration phenomena such as elliptic relations or combination of synonymy and derivation relations (e.g. heat and thermal) where the categories may be different.

Fuzzier relations such as the traditional see-also relations of terminologies are also very useful. Once a link is established between two terms, it is sometimes easy to interpret for the terminology users. Moreover, for applications such as document retrieval, the link itself is often more important than its very type.

Synonymy We use a synonymy definition close to that of WordNet (Miller et al., 1993). It is defined as an equivalence relation between terms having the same meaning in a particular context. The transitivity rule cannot be applied to the links extracted from the dictionary. Indeed, while the synonymy is sometimes very contextual in the dictionary, the links appear in the data without context information and would produce a great deal of errors. Thus, for instance, the synonymy links between the adjectives *polaire* (polar) and *glacial* (icy) and the adjectives *glacial* (cold) and *insensible* (insensitive) would allow to deduce a wrong synonym link between *polaire* and *insensible*.

Moreover, tests carried out on dictionary samples show that the relevant links which

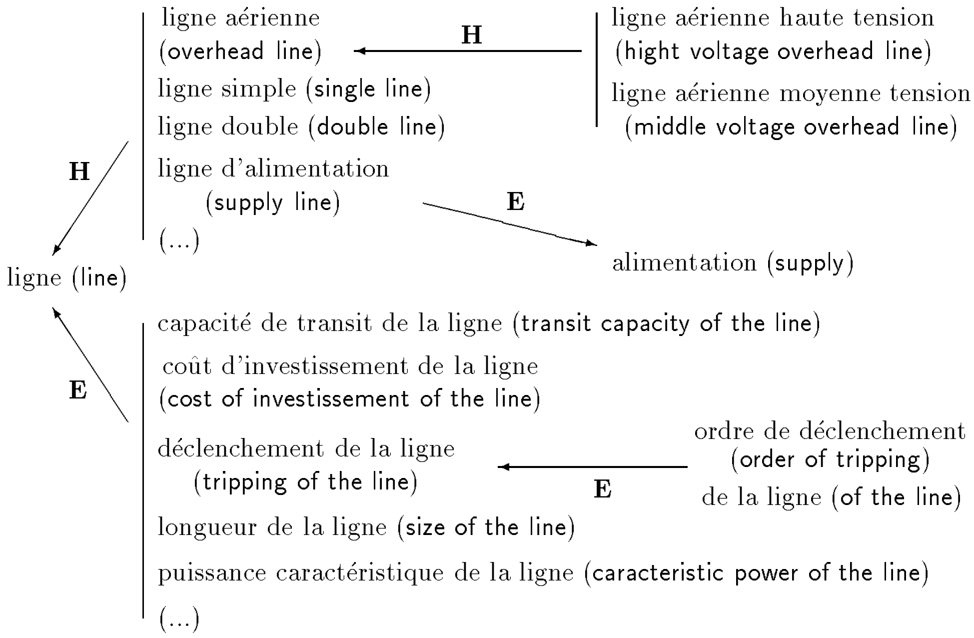

Figure 3: Fragment of the syntactic network (H = head, E = expansion).

	Nouns	Adjectives	Total
Number of simple terms extracted	2 831	597	3 428
Number of retained words at the filtering step	1 134	408	1 542
Percentage of retained words at the filtering step	40%	68%	45%

Table 1: Coverage of the corpus by the dictionary.

could be added thanks to the transitivity rules already exist in the dictionary. For instance the following words are synonymous pairwise: *logement* (accommodation), *demeure* (residence), *domicile* (residence) and *habitation* (house).

We consider all links provided by the dictionary as expressing synonymy relation between simple candidate terms and design a two-step automatic method to infer links between complex candidate terms.

2.2 First step: Dictionary data filtering

In order to reduce the database, we first filter the relevant dictionary links for the studied document. For instance, the link *matériel* (equipment) / *équipement* (fittings) is selected because its both ends, *matériel* and *équipement* exist in the studied corpus. For this document, 3 369 synonymy links between 1 542 simple terms are preserved.

Table 1 shows the results of the filtering step in regard to the coverage of our corpus by the dictionary.

2.3 Second step: Detection of synonymous candidate terms

Assuming that the semantics and the synonymy of the complex candidate terms are compositional, we design three rules to detect synonymy relations between candidate terms. Considering two candidates terms, if one of the following conditions is met, a synonymy link is added to the terminological network:

- the heads are identical and the expansions are synonymous (*collecteur général* (general collector) / *collecteur commun* (common collector));
- the heads are synonymous and the expansions are identical (*matériel électrique* (electric equipment) / *équipement électrique* (electrical fittings));
- the heads are synonymous and the expansions are synonymous (*marche normale* (normal running) / *bon fonctionnement* (right working));

We first use the dictionary links as a bootstrap to detect synonymy links between complex candidate terms. Then, we iterate the process by including the newly detected links in our base until no new link can be found. In the present experiment, the process ends up after three iterations.

3 Results and study of the detected links

3.1 Various detected links

Synonymy links 396 links between complex candidate terms (i.e. noun phrases) are inferred by this method. An expert of the domain validated 37% of them (i.e. 146 links, cf. table 2) as real synonymy links: *hauteur d'eau* (water height) / niveau d'eau (level of water), détérioration notable (notable deterioration) / dégradation importante (important damage) (cf. figure 4).

	Number	Percentage
Validated links	146	37%
Unvalidated links	250	63%
Total	396	100%

Table 2: Results of the link validation.

Most of the synonymy links between candidate terms are detected at the first iteration (383 liens out of 396). The majority of the validated links are given by the two first rules: 89 validated links out of 206 with the first rule (*admission d'air* (air intake) / *entrée d'air* (air entry)), 49 out of 105 with the second (*toit flottant* (floating roof) / *toit mobile* (movable roof) and *collecteur général* (general collector) / *collecteur commun* (common collector)). Obviously, the last rule has a lower precision rate: 8 out of 85 (*fausse manœuvre* (wrong operation) / *mauvaise manipulation* (bad handling)). However, it infers important links which are difficult to detect by hand.

Other useful links On the whole, the expert judged that half of the detected links are useful for the terminology structuration even if he rejected some of them as real synonymy links (cf. figure 5). Our method detects different types of links: meronymy, antonymy, relations between close concepts, connected parts of a whole mechanism, etc.

The meronymy links are the most numerous after synonymy (*rapport de sûreté* (safety report) / *analyse de sûreté* (safety analysis)). In the previous example, whereas *rapport* (report) and *analyse* (analysis) are given as synonyms by the general language dictionary (which is context-free), their technical meanings in our document are more specific. Therefore, *rapport de sûreté* is a meronym rather than a synonym of *analyse de sûreté* in the studied document.

Other detected links allow to group the candidate terms which refer to related concepts. For instance, we detected a link between the device *ligne de vidange* (draining line) and the place *point de purge* (blow-down point) which is relevant since a draining line ends at a blow-down point. Likewise, it is useful to link *fin de vidange* (draining end) which designates an operation and *destination des purges* (blow-down destination) which is the corresponding equipment.

The expert considered that the link between the candidate terms (*commande mécanique* (mechanical control) / *ordre automatique* (automatic order)) expresses an antonymy relation, although it is infered from the synonymy relation of the dictionary *mécanique* (mechanical) / *automatique* (automatic). It appears that those adjectives have a particular meaning in the present corpus. Therefore, every link detected from this "synonymy" link is an antonymy one.

Those links express various relations sometimes difficult to name, even by the expert. Such links are important in a terminology.

3.2 Polysemy, elision and metaphor

Most real errors are due to the lack of context information for polysemic words and the noisy data existing in the dictionary. For instance the French word *temps* means either time or weather. According to the dictionary,*temps* (weather) is a synonym of *temperature* (temperature)[2], but this meaning is excluded from the present corpus. Since we cannot distinguish the different meanings, the synonymy of *temps* / time and temperature is taken for granted. *Temps attendu* (expected time) and *température attendue* (expected tempera-

[2] It would be more precise to interpret it as analogous words.

Term 1	Term 2
détérioration notable (notable deterioration)	dégradation importante (important damage)
fausse manoeuvre (wrong operation)	mauvaise manipulation (bad handling)
action de l'opérateur (action of the operator)	intervention de l'opérateur (intervention of the operator)
capacité interne (internal capacity)	volume interne (internal volume)
capacité totale (total capacity)	volume total (total volume)
capacité utile (useful capacity)	volume utile (useful volume)
limite de solubilité (limit of solubility)	seuil de solubilité (solubility threshold)
marche manuelle (manual running)	fonctionnement manuel (manual working)
tests périodiques (periodic tests)	essais périodiques (periodic trials)
hauteur d'eau (water height)	niveau d'eau (level of water)
panneau de commande (control panel)	tableau de commande (control board)

Figure 4: Examples of synonymy links between complex candidate terms.

Term 1	Term 2
essai en usine (test in plant)	expérience d'exploitation (experiment of exploitation)
ligne de vidange (draining line)	point de purge (blow-down point)
fonction d'un temps (fonction of a time)	effet d'une température (effect of a temperature)
froid normal (normal cold)	refroidissement correct (correct cooling)
rapport de sûreté (safety report)	analyse de sûreté (safety analysis)
solution d'acide borique (solution of boric acid)	dissolution de l'acide borique (dissolving of the boric acid)
température attendue (expected temperature)	temps attendu (expected time)
température normale (normal temperature)	temps normal (normal time)
organes de commande (control devices)	organes d'ordre (order devices)
gros débit (big flow)	plein débit (full flow)
activité importante (important activity)	activité élevée (high activity)
commande mécanique (mechanical control)	ordre automatique (automatic order)
risques de corrosion (risk of corrosion)	risques de destruction (risk of destruction)

Figure 5: Examples of rejected links between complex candidate terms.

ture) are thus given as synonymous. This type of wrong links is rather important in the list presented to the expert: between 10 to 20 links out of 396.

On the contrary, about ten wrong links are due to the elision of common terms in the domain. For instance, the term *activité* (activity) which actually corresponds to the term *radioactivité* (radioactivity) in the document is given as a synonym of *énergie* (energy) in the dictionary. We have detected links such as *activité haute* (high activity) / *haute énergie* (high energy).

As regards metaphor, we have observed that it preserves semantic relation. For instance, in graph theory, the link (*arbre* (tree) / *feuille* (leaf)) can be inferred from the meronyny information of general dictionary.

Those types of wrong links are easily identified during the validation. Some exceptions rules can be designed to first regroup those links and then eliminate them. With that aim, we plan to use dictionary definitions.

3.3 Evaluation

The inferred links express not only synonymy, but also other relations which may be difficult to name. Apart from real errors, these fuzzy see-also relations are useful in the context of a consultation system.

The results of this first experiment are encouraging. Although the precision rate and the number of links are low (37%, 396 links), the use of complementary methods (e.g. detection of syntactic variants) would allow to propagate these links and increase their number. Also, the use of other knowledge sources or different methods (Habert et al., 1996) is necessary to increase precision rate and find links between more technical candidate terms.

As regards the improvement of such a method, the terminology acquisition by an expert will take tens of hours while the automatic extraction takes one hour and the validation of the links has been done in two hours.

The main difficulty is to evaluate the recall in the results because there is no standard reference in that matter, giving the overall relevant relations in the document. One may think that the comparison with links manually detected by an expert is the best evaluation, but such manual detection is subjective. Regarding the validation by several experts, it is well-known that such validation would give different results depending on the background of each expert (Szpakowicz et al., 1996). So, we are reduced to compare our results with those obtained by different methods even though they are not perfect either. We are planning to compare the clusters found by our method with the clustering one of (Assadi, 1997) to study how the results overlap and are complementary.

4 Related works

The variant detection in specialized corpora must be taken into account for information retrieval. This complex operation involves the semantic as well as the morphological and syntactic level. (Jacquemin, 1996) design a unification-based partial parser FASTER which analyses raw technical text while meta-rules detect morpho-syntactic variants of controlled terms (*blood cell*, *blood mononuclear cell*). By using morphological and part-of-speech modules, the system are extended to the verbal phrases (*tree cutting*, *tree have been cut down*) (Klavans et al., 1997). Dealing with syntactic paraphrase in the general language, (Dras, 1997) propose a similar representation by using the STAG formalism to detect syntactic related sentences. Because we deal with the semantic level, our work is complementary of those.

Semantic variation is rarely studied in specialized domains. Works on word similarity and word sense disambiguation are generally based on statistical methods designed for large or even very large corpora (Hindle, 1990; Agirre and Rigau, 1996). Therefore, they cannot be applied for technical documents which usually are medium size corpora. However, dealing with already linguistic filtered data, (Assadi, 1997) aims at statistically build rough clusters supposing that similar candidate terms have similar expansions. Then he relies on human expertise for the semantic interpretation. It differs from our work which tries to automatically explicit the semantic relations. In order to disambiguate noun objects in a short text (30 000 words), (Li et al., 1995) design heuristic rules using semantic similarity information in WordNet and verbs as context. Their system disambiguate an encouraging number on noun-verb pairs if one considers single and multiple sense assigned to a word.

In (Basili et al., 1997), the lexical knowledge base WordNet (Miller et al., 1993) is used as a bootstrap for verb disambiguation. They tune it to the domain of the studied document by taking into account the contexts in which the verbs are used. This tuning leads both to eliminate certain semantic categories and to add new ones. For instance, the category *contact* is created for the verb *to record*. The resulted sense classification is thus a better description of the verb specialized meanings.

Our symbolic and dictionary-based approach is close that of (Basili et al., 1997). They both use general language information (traditional dictionary vs. WordNet) for specialized corpora. However, their goals differ: disambiguation vs. semantic relation identification.

5 Conclusion and future works

The use of a synonym dictionary and the rules of synonymous candidate terms detection we have designed allow to extract an encouraging number of links in a very technical corpus. An expert validated these links. More than one third of the detected links are synonymy relations. Beside synonymy, our method detects various kinds of semantic variants. Wrong links due to the polysemy can be easily eliminated with exception rules by comparing selectional patterns and generalized contexts (Basili et al., 1993; Grishman and Sterling, 1994).

Our work shows that general semantic data are useful for the terminology structuration and the synonym detection in a corpus of specialized language. The results show that semantic variants can be automatically detected. Of course, the number of acquired links is relatively low but our method is not to be used in isolation.

Acknowledgment

This work is the result of a collaboration with the Direction des Etudes et Recherche (DER) d'Electricité de France (EDF). We thank Marie-Luce Picard from EDF and Benoît Habert from ENS Fontenay-St Cloud for their help, Didier Bourigault and Jean-Yves Hamon from the Institut de la Langue Française (INaLF) for the dictionary and Henry Boccon-Gibod for the validation of the results.

References

E. Agirre and G. Rigau. 1996. Word sense disambiguation using conceptual density. In *Proceedings of COLING'96*, pages 16–22, Copenhagen, Danmark.

H. Assadi. 1997. Knowledge acquisition from texts: Using an automatic clustering method based on noun-modifier relationship. In *Proceedings of ACL'97 - Student Session*, Madrid, Spain.

Roberto Basili, Maria Teresa Pazienza, and Paola Velardi. 1993. Acquisition of selectional patterns in sublanguages. *Machine Translation*, 8:175–201.

Roberto Basili, Michelangelo Della Rocca, and Maria Teresa Pazienza. 1997. Contextual word sense tunig and disambiguation. *Applied Artificial Intelligence*, 11:235–262.

D. Bourigault. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of COLING'92*, pages 977–981, Nantes, France.

Mark Dras. 1997. Representing paraphrases using synchronous tree adjoining grammars. In *proceedings of the 1997 Australian NLP Summer Workshop*, Syndney, Australia.

Ralph Grishman and John Sterling. 1994. Generalizing automatically generated selectional patterns. In *Proceedings of Coling'94*, volume 3, pages 742–747, Kyoto.

C. Gros, H. Assadi, N. Aussenac-Gilles, and A. Courcelle. 1996. Task models for technical documentation accessing. In *Proceedings of EKAW'96*, Nottingham.

Benoît Habert, Elie Naulleau, and Adeline Nazarenko. 1996. Symbolic word clustering for medium-size corpora. In *Proceedings of COLING'96*, volume 1, pages 490–495, Copenhagen, Danmark, August.

D. Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of ACL'90*, pages 268–275, Pittsburgh, PA.

C. Jacquemin. 1996. A symbolic and surgical acquisition of terms through variation. In E. Riloff et G. Scheler S. Wermter, editor, *Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing*, pages 425–438, Springer.

J. Klavans, C. Jacquemin, and E. Tzoukermann. 1997. A natural language approach to multi-word term conflation. In *Proceedings of the third Delos Workshop – Cross-Language Information Retrieval*.

Xiaobin Li, Stan Szpakowicz, and Stan Matwin. 1995. WordNet-based algorithm word sense disambiguation. In *Proceedings of IJCAI-95*, pages 1368–1374, Montreal, Canada.

G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller. 1993. Introduction to WordNet: An on-line lexical database. Technical Report CSL Report 43, Cognitive Science Laboratory, Princeton.

Stan Szpakowicz, Stan Matwin, and Ken Barker. 1996. WordNet-based word sense disambiguation that works for short texts. Technical Report TR-96-03, Department of Computer Science, University of Ottawa, Ontario, Canada.