Using general semantic information to help the terminology structuration

Thierry Hamon & Adeline Nazarenko
Laboratoire d'Informatique de Paris-Nord
Université Paris-Nord
93430 Villetaneuse, FRANCE
[*firstname.lastname*@lipn.univ-paris13.fr]

Abstract

We report in this paper the results of a preliminairy experience on the relevance of general semantic information in the context of a terminology structuration. As resources supplying specialised semantic information in the domain of the studied documents are missing, we use general semantic information to structure a terminology. By exploiting a general synonymy dictionary, we attempt to infer synonymy links between candidate terms extracted from French technical document. The three rules designed allow us to propose synonymy links between complex candidate terms to an expert of the domain. An interesting part of the detected links have been validated as synonymy links. On the whole, it appeared that half of the overall links is useful. We have noted also that numerous errors are due to the polysemy of general language words, coarse-grained distinction of some meanings and the elision of words widely used in the domain. Thus, we argue that general semantic information must be used with others resources such as morpho-syntactic links and is a useful aid for the terminology structuration.

Introduction

Using a general language dictionary to structure a terminology

Terminologies are essential in specialised domains since they structure terms into a semantic network by providing hierarchical (*is_a*) and equivalence (*synonymy* and *see_also*) links between them (cf. figure 1). However, they are seldom available for specific technical documents and thus, they must be built. In the context of Technical Documentation Consulting System (Gros et al., 1996), such a terminology and especially synonymy links allow to index terms with a single word entry.

In the perspective of the construction of a terminology from technical documents, several tools are developed to help terminologists to extract and structure terms semantically. The method can be viewed as a two-step process. The candidate terms are first automatically extracted from the document. Then, the list of candidate terms is structured into a semantic network. The experiment reported here focuses on the second step.

As semantic information in specialised domains is seldom available, we attempt to use general semantic information extracted from a machine-readable dictionary to help term structuration, and add semantic links between terms extracted from a document. Therefore, we evaluate the relevance and usefulness of general semantic resources for the detection of synonymy between complex 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 distributed among the different senses (cf. figure 2) of each word entry. It is exploited as a synonym dictionary.

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.

A ***term*** is defined as a simple or complex lexical unit pointing out an accurate concept in a technical document, i.e. a terminological unit (Bourigault, 1992). Therefore, a ***candidate term*** is a lexical unit which can be a term if a domain expert validates it.

General dictionaries in corpora

The use of general language resources in specialised documents is rarely studied from the NLP point of view. However, dealing with the problem of domain identification in document, (Haas, 1992) evaluates how helpful general language dictionaries can be. Studying of the coverage of general and specialized dictionaries for technical documents, she argues that general dictionaries are useful but not sufficient.

On the other hand, specialized dictionaries are difficult to use in an automatic process because they are not structured and because it is difficult to choose the dictionary whose coverage fits the best the studied document without trying every dictionary. She concludes on the complementary of the both kinds of resources, giving priority to the specialized dictionary information.

General language resources are generally used in methods dealing with word similarity and word sense disambiguation in general language corpora. (Yarowsky, 1992) used the Roget's thesaurus with a mutual-information-like estimate to disambiguate words and (Resnik, 1995) propose a measure of semantic similarity based on the WordNet taxonomy (Miller et al., 1993). Corpus-based methods (Dagan et al., 1994; Hindle, 1990) use cooccurence probability to determine the similarity between words. However, these

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***.

types of methods rely on large and very large corpora.
As we use small size corpora, the methods described above cannot be applied. We adopted a dictionary- and rule-based approach which differs from those cited above because it deals with specialized texts. In that respect, our approach is close to (Basili et al., 1997).

Experimental data

The experiment have been carried out on three technical documents dealing respectively with electric power plant, electric network planning and software engineering. The proportion of dictionary words appearing in these documents is respectively 45 %, 57% and 65%. We presented the results on the first document. On the basis of the dictionary coverage, this document can be considered as the more technical document and as the worse case for our experiment.
This corpus has been analysed by the terminology extraction software LEXTER (Bourigault, 1992) which has extracted 12 043 words, that is 2 831 nouns, 597 adjectives and 8 615 noun phrases. Each complex candidate term (*matériel électrique*, **electric equipment**) is analysed into an head (*matériel*, **equipment**) and an expansion (*électrique*, **electric**) and they form a syntactic network (cf. figure 3).

Term Synonymy detection

Synonymy defined as an equivalence

We define the synonymy relation as an equivalence relation between terms having the same meaning in a particular context. This definition is close to that of WordNet (Miller et al., 1993) where the synonymy is defined in a given linguistic context.
Contextuality is important here. Although the synonymy relation is transitive, we cannot apply transitivity to the context-free dictionary links.
Actually, applying transitivity as in (Warnesson, 1985) would give wrong synonymy links such as a synonymy relation between *polaire* and *insensible* which would be inferred from the links between the adjectives *polaire* (**polar**) and *glacial* (**icy**) and the adjectives *glacial* (**cold**) and *insensible* (**insensitive**).
The reason is that, although synonymy links have been built with a certain context in the dictionary, this context is implicit. Dictionary users can be easily found the implicit information and they know when the transitivity rule can be used, but this contextualization process is difficult to simulate automatically.
However, we carried out tests on dictionary samples showing that the relevant links which could be added thanks to the transitivity rule, already exist in the dictionary. For instance the following words are synonymous pairwise: *logement* (**accomodation**), *demeure* (**residence**), *domicile* (**residence**) and *habitation* (**house**).

Methods

Dictionary links 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 *equipement* exist in the studied corpus. For this document, 3 369 synonymy links are preserved.
Table 1 shows the results of the filtering step with respect to the dictionary coverage.

	Nouns	Adjectives	Total
Number of words in the base	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: Dictionary coverage of the studied document.

Detection of synonymous candidate terms Assuming that the semantics and the synonymy of the complex candidate terms are compositional, we have designed three rules to infer 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**);

The method is an iterative process where the dictionary links are considered as synonymy links between simple candidate terms (generally composed of a single word). We first used these links as a bootstrap to infer synonymy links

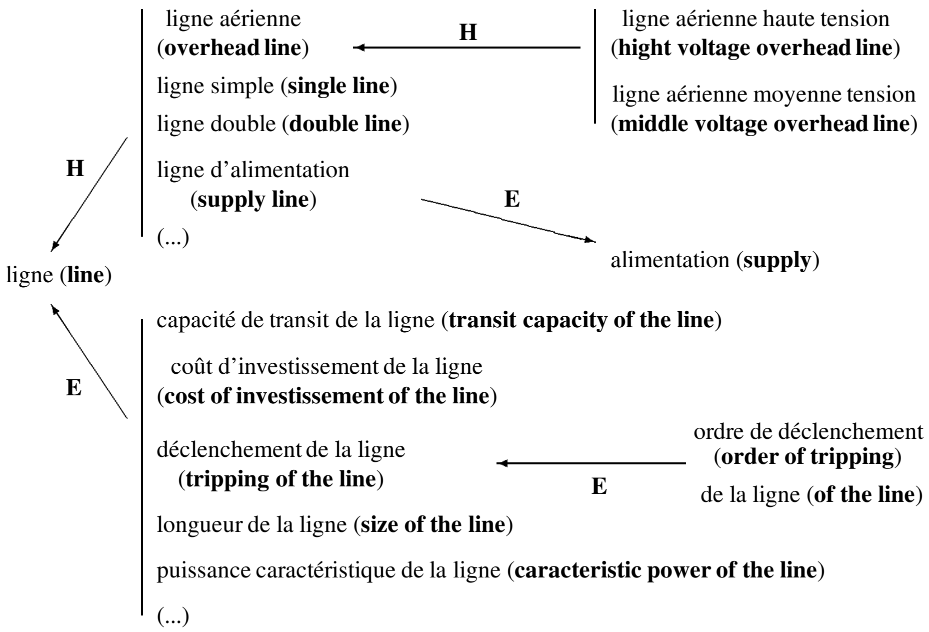

Figure 3: *Fragment of the terminological network provided by LEXTER (H = head, E = expansion).*

between complex and multi-word candidate terms, using the three rules presented above. Then, the newly inferred links are included in the base, and the process is iterated until no new link can be found. In the reported experiment, the process ended up after three steps.

Results and study of the detected links

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 by the expert.

Most of the synonymy links between candidate terms are inferred at the first iteration (383 links out of 396).

The majority of the validated links are given by the two first rules in the same proportion: 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 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. Thus, 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 considered as synonyms in the general language and in the 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 inferred from the synonymy relation of the dictionary *mécanique* (**mechanical**) / *automatique* (**automatic**). It appears that those adjectives have a specific meaning in the present corpus: a mechanical control needs the operator's intervention. Therefore, every link inferred from this "synonymy" link is an antonymy one.

The figure 5 presents a sample of the rejected links.

Polysemy, elision and metaphor

Most real errors are due to the lack of context information for words which are polysemic in the general language. We

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 candidate terms which are validated by the expert.

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 links between candidate terms which have been rejected by the expert as synonymy links.

temps : $<11>$ moment (**while**), période (**period**),
(**time**/*weather* $<121>$ date (**date**), époque (**period**), passé (**past**), présent (**present**)
$<122>$ âge (**age**), génération (**generation**) , siècle (**century**)
$<13>$ conjugaison (**conjugaison**)
$<14>$ durée (**length of time**)
$<2>$ air (**air**), température (**temperature**), vent (**wind**)
intempérie (**bad weather**), orage (**storm**)

Figure 6: *Temps* entry in the dictionary ***Le Robert***.

have used for this experiment a synonym dictionary designed for human users and not for a computational task. Moreover, the meanings of each headword are more or less accurate. While a human user is able to infer the context from a list of coarse-grained distinct synonyms (cf. figure 6), the lack of context generates noise in the list of inferred links when dictionaries are used in a computational task.

For instance the French word *temps* means either **time** or **weather**. With the meaning of **weather**, it is synonymous with *temperature* (**temperature**), 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 temperature**) 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**).

Evaluation

The results of this first experiment are encouraging. The inferred links express not only synonymy, but also other relations which may be difficult to name. Apart from real errors, some rejected links can be viewed as expressing the see_also relation of terminologies. How fuzzy these relations are, they are useful in the context of a consultation system.

Although the number of links are low (396 links), the use of complementary methods (e.g. detection of morpho-syntactic links) and the syntactic network would allow to propagate these links and increase their number. Also, other knowledge sources or different methods (Habert et al., 1996) are necessary to find links between more technical candidate terms.

The main difficulty is to evaluate the silence in the results, because there is no standard reference in that matter. One may think that the comparison with links manually detected by an expert is the best evaluation, but such manual detection is time-consuming and subjective. 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 our method with the clustering one of (Assadi, 1997).

Although these first results are mixed, our work shows that the semantic data of the general language can be a useful aid for the terminology structuration and the synonym search in technical documents. First of all, the automatically detected links and those added by the expert are generally different. The experts are not used to detect such links in a document : they rather focus on specialized vocabulary and synonymy. Both approaches therefore seems complementary. General semantic information and the list of synonymous terms can also be helpful for the terminology control process. In traditional terminological work, this process consists, if ever several synonymous terms can be found, in choosing one of them as the controlled or preferred term which will be a terminological entry, others are referred as rejected terms whose use is discouraged.

Conclusion

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. They are validated by an expert of the domain. It appears that more than one third of the links are strict synonymy relations between complex terms. If we consider a wide synonymy relation i.e. including the see_also relation, half of the detected links are useful for the terminology structuration.

However, the low number of inferred links points out that our method should not be used in isolation and must be combined with other knowledge and methods. In particular, the links can be used to generalize contexts as in (Basili et al., 1993; Grishman and Sterling, 1994) and thus serve as a bootstrap for methods which rely on contexts to bring out more specialized semantic links.

Acknowledgements

This work has been taken place at the Direction des Etudes et Recherches (DER) d'Electricité de France (EDF) during an first author's internship. We would like to thank Cécile Gros and Marie-Luce Picard from EDF, Benoît Habert from the ENS Fontenay-St Cloud for their help, Didier Bourigault and Jean-Yves Hamon from the Institut National de la Langue Française (INaLF) for the dictionary used in this work, and Henry Boccon-Gibod for the validation of the results.

References

Assadi, Houssem (1997). Knowledge acquisition from Texts: Using an automatic Clustering Method Based on Noun-Modifier Relationship. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistic (ACL'97) - Student Session*. Madrid, Spain.

Basili, Roberto, Pazienza, Maria Teresa & Velardi, Paola (1993). Acquisition of Selectional Patterns in Sublanguages. *Machine Translation*, vol. 8, 175–201.

Basili, Roberto, Rocca, Michelangelo Della & Pazienza, Maria Teresa (1997). Contextual Word Sense Tunig and Disambiguation. *Applied Artificial Intelligence*, vol. 11, 235–262.

Bourigault, Didier (1992). Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases. In *Proceedings of COLING-92* (pp. 977–981). Nantes, France.

Dagan, Ido, Pereira, Fernando & Lee, Lillian (June 1994). Similarity-Based Estimation of Word Cooccurrence Probabilities. In *Proceedings of the 32nd Annual meeting of the Association for Computational Linguistics (ACL'94)* (pp. 272–278). New Mexico State University, New Mexico, USA.

Grishman, Ralph & Sterling, John (1994). Acquisition of selectional patterns. In *Proceedings of COLING-92* (pp. 658–664). Kyoto, Japan.

Gros, Cécile, Assadi, Houssem, Aussenac-Gilles, Nathalie & Courcelle, Arnaud (May 1996). Task Models for Technical Documentation Accessing. In *Proceedings of EKAW'96*. Nottingham.

Haas, Stephanie W. (1992). Covering the vocabulary of technical abstracts using standard and specialized dictionaries. *Journal of Information Science*, vol. 18, 363–373.

Habert, Benoît, Naulleau, Elie & Nazarenko, Adeline (August 1996). Symbolic word clustering for medium-size corpora. In *Proceedings of the 16th International Conference on Computational Linguistics* (pp. 490–495). Copenhagen, Danmark.

Hindle, Donald (1990). Noun Classification from Predicate-Argument Structures. In *Proceedings of the 28st Annual meeting of the Association for Computational Linguistics (ACL'90)* (pp. 268–275). Pittsburgh, Pennsylvania.

Miller, George A., Beckwith, Richard, Fellbaum, Christiane, Gross, Derek & Miller, Katherine (1993). *Introduction to WordNet: An On-line Lexical Database*. Technical Report CSL Report 43. Cognitive Science Laboratory, Princeton.

Resnik, Philip (1995). Disambiguating Noun Groupings with Respect to WordNet Senses. In *Proceedings of the Third Workshop on Very Large Corpora*. MIT, USA.

Warnesson, Isabelle (1985). Applied Linguistics: Optimization of Semantic Relations by Data Aggregation Techniques. *Applied Stochastic Models and Data Analysis*, vol. 1, 121–141.

Yarowsky, David (August 1992). Word-Sense Disambiguation Using Statistical Models of Roget's Cateories Trained on Large Corpora. In *Proceedings of COLING-92* (pp. 454 – 460). Nantes, France.