Symbolic word clustering for medium-size corpora

Benoît Habert* and Elie Naulleau* ** and Adeline Nazarenko*
*Equipe de Linguistique Informatique
Ecole Normale Supérieure de Fontenay-St Cloud
31 av. Lombart, F-92260 Fontenay-aux-Roses
Firstname.Name@ens-fcl.fr
**Direction des Etudes et Recherches - Electricité de France
1, av. du G^{al} de Gaulle, F-92141 Clamart
Firstname.Name@der.edfgdf.fr

Abstract

When trying to identify essential concepts and relationships in a medium-size corpus, it is not always possible to rely on statistical methods, as the frequencies are too low. We present an alternative method, symbolic, based on the simplification of parse trees. We discuss the results on nominal phrases of two technical corpora, analyzed by two different robust parsers used for terminology updating in an industrial company. We compare our results with Hindle's scores of similarity.

Subjects Clustering, ontology development, robust parsing, knowledge acquisition from corpora, computational terminology

1 Identifying word classes in medium-size corpora

In companies with a wide range of activities, such as EDF, the French electricity company, the rapid evolution of technical domains, the huge amount of textual data involved, its variation in length and style imply building or updating numerous terminologies as NLP resources. In this context, terminology acquisition is defined as a twofold process. On one hand, a terminologist must identify the essential entities of the domain and their relationships, that is its ontology. On the other hand, (s)he must relate these entities and relationships to their linguistic realizations, so as to isolate the lexical entries to be considered as certified terms for the domain.

In this paper, we concentrate on the first issue. Automatic exploration of a sublanguage corpus constitutes a first step towards identifying the semantic classes and relationships which are relevant for this sublanguage.

In the past five years, important research on the automatic acquisition of word classes based on lexical distribution has been published (Church and Hanks, 1990; Hindle, 1990; Smadja, 1993; Grefenstette, 1994; Grishman and Sterling, 1994). Most of these approaches, however, need large or even very large corpora in order for word classes to be discovered[1] whereas it is often the case that the data to be processed are insufficient to provide reliable lexical information. In other words, it is not always possible to resort to statistical methods. On the other hand, medium size corpora (between 100,000 and 500,000 words: typically a reference manual) are already too complex and too long to rely on reading only, even with concordances. For this range of corpora, a pure symbolic approach, which recycles and simplifies analyses produced by robust parsers in order to classify words, offers a viable alternative to statistical methods. We present this approach in section 2. Section 3 describes the results on two technical corpora with two different robust parsers. Section 4 compares our results to Hindle's ones (Hindle, 1990).

2 Simplifying parse trees to classify words

2.1 The need for normalized syntactic contexts

As Hindle's work proves it, among others (Grishman and Sterling, 1994; Grefenstette, 1994), the mere existence of robust syntactic parsers makes it possible to parse large corpora in order to automate the discovery of syntactic patterns in the spirit of Harris's distributional hypothesis. However, Harris' methodology implies also to simplify and transform each parse tree[2] , so as to obtain so-called "elementary sentences" exhibiting the main conceptual classes for the domain (Sager

[1] For instance, Hindle (Hindle, 1990) needs a six million word corpus in order to extract noun similarities from predicate-argument structures.

[2] Changing passive into active sentences, using a verb instead of a nominalization, and so on.

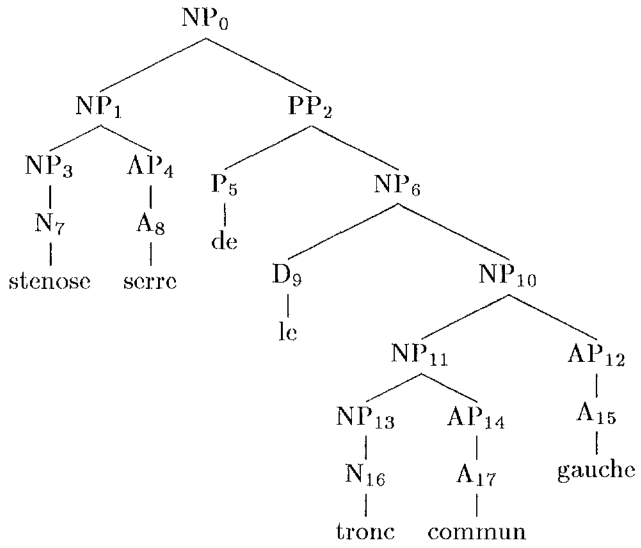

Figure 1: Parse tree for *stenose serre de le tronc commun gauche*

et al., 1987).

In order to automate this normalization, we propose to post-process parse trees so as to emphasize the dependency relationships among the content words and to infer semantic classes. Our approach can be opposed to the a prior one which consists in building simplified representations while parsing (Basili et al., 1994; Metzler and Haas, 1989; Smeaton and Sheridan, 1991).

2.2 Recycling the results of robust parsers

For the sake of reusability, we chose to add a generic post-processing treatment to the results of robust parsers. It implies to transduce the trees resulting from different parsers to a common format.

We experimented so far two parsers: AlethGram and Lexter, which are being used at DER-EDF for terminology acquisition and updating. They both analyze corpora of arbitrary length. AlethGram has been developped winthin the GRAAL project[3]. LEXTER has been developped at DER-EDF (Bourigault, 1993). In this experiment, we focussed on noun phrases, as they are central in most terminologies.

2.3 The simplification algorithm

The objective is then to reduce automatically the numerous and complex nominal phrases provided by AlethGram and LEXTER to elementary trees, which more readily exhibit the fundamental binary relations , and to classify words with respect to these simplified trees.

For instance, from the parse tree for *stenose serre de le tronc commun gauche*[4] (cf. fig. 2, in which non terminal nodes are indexed for reference purposes), the algorithm[5] yields the set of elementary trees of figure 1. The trees *a* and *c* correspond to contiguous words in the original sequence, whereas *b* and *d* only appear after modifier removal (see below).

Two types of simplifications are applied when possible to a given tree:

1. *Splitting*: Each sub-tree immediately dominated by the root is extracted and possibly further simplified. For instance, removing node NP_0 yields two sub-trees: NP_1, which is elementary (see below) and PP_2, which needs further simplification.

2. *Modifier removal*: Within the whole tree, every phrase which represents a modified constituent is replaced by the corresponding non modified constituent. For example, in NP_0, the adjectival modifier *serre* is removed, as well as the determiner and the adjectives

[3] The Eureka GRAAL project gathers in France GCI-ERLI (prime contractor), EDF, Aerospatiale and Renault.

[4] *Tight stenosis of left common mainstem.* In both parsers, the accents are removed during the analysis, the lemmas are used instead of inflected forms. Additionally, for simplification purposes, a contracted word like *du* is considered as a *preposition- determiner* sequence.

[5] See (Habert et al., 1995) for a detailled presentation. The corresponding software, SYCLADE, has been developped by the first author.

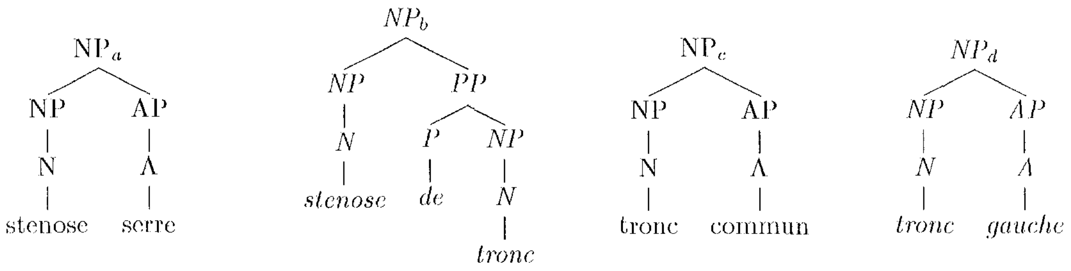

Figure 2: Elementary trees for *stenose serre de le tronc commun gauche*

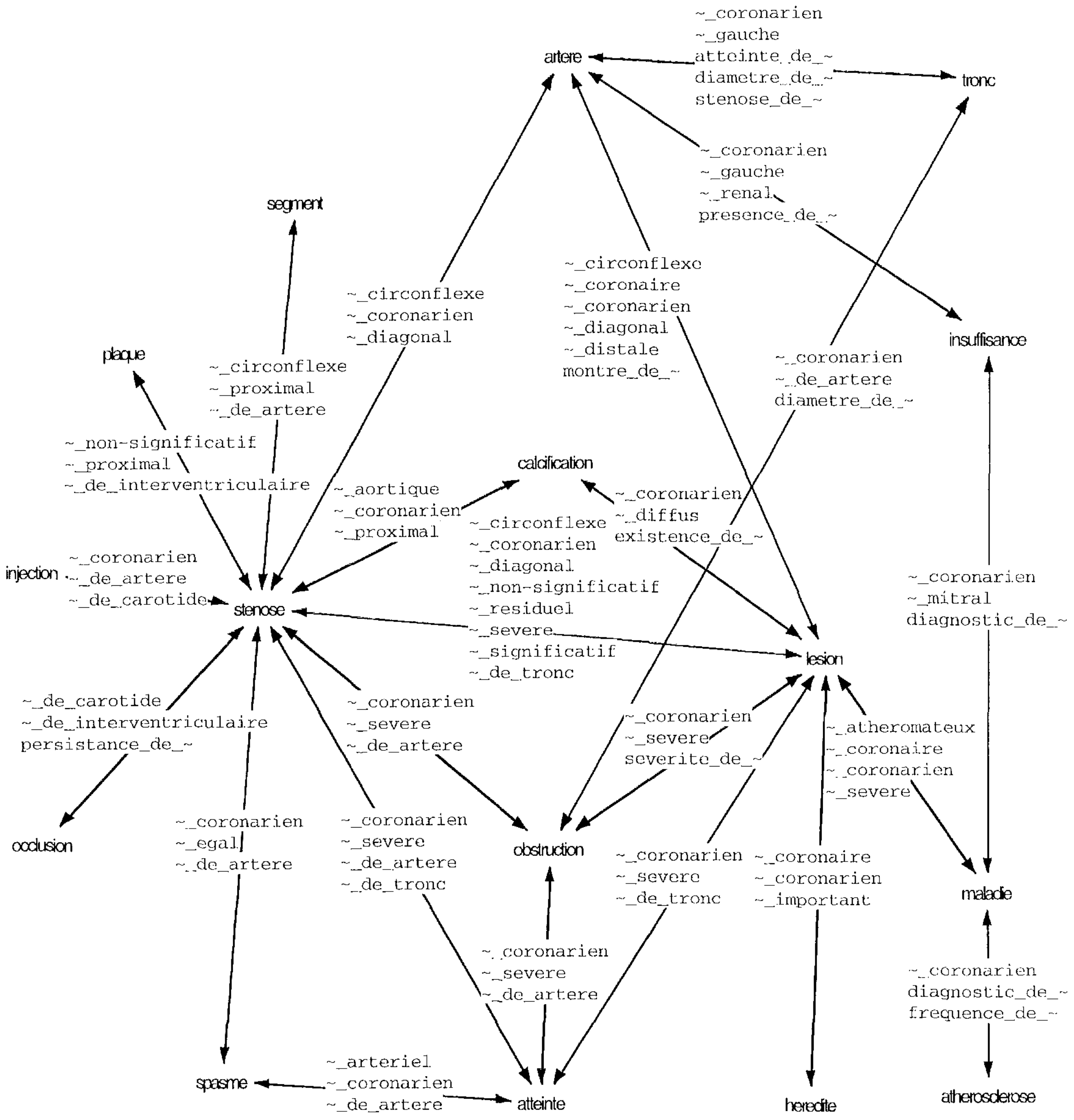

Figure 3: Example of a strongly connected component (CMC corpus)

modifying *tronc*, which leads to elementary tree *b*.

When the current tree is *elementary*, the simplification process stops. Before processing the set of original parse trees, one must declare the trees which must not be simplified any further. In this experiment, are considered as elementary the nominal trees which exhibit a binary relation between two "content" words, for instance between two N in an N P N sequence.

2.4 From elementary contexts to word classes

The resulting collocations are controlled by the syntactic relationships structuring the parse trees, which is not the case for window-based approaches (Church and Hanks, 1990), even when they use part-of-speech labels (Smadja, 1993; Daille, 1994). In the example, *gauche* is not related to *stenose*, as it does not modify this noun.

The elementary trees lead to classes of syntactic contexts. For instance, from the tree corresponding to *stenose serre*, two classes of contexts are created. The first one, *stenose* ~, in which ~ stands for the pivot word, contains *serre*, whereas the second one, ~ *serre*, contains *stenose*. At the end of the simplification process, these classes have been completed and other ones created. We claim that the semantic similarity between two lexical entries is in proportion with the number of shared contexts. For instance, in one of our corpora, *stenose* shares 8 contexts with *lesion*.

In order to get a global vision of the similarities relying on elementary contexts, a graph is computed. The words constitute the nodes. A link corresponds to a certain number of shared contexts (according to a chosen threshold). The edges are labelled with the shared contexts. The strongly connected components[6] and the cliques[7] are computed as well, as they are the most relevant parts of the graph, on topological grounds. The underlying intuition is that a connected component relates neighboring words (Bensch and Savitch, 1995) and that the cliques tend to isolate similarity classes. An extract of a connected component, with 3 as a threshold, appears in figure 3.

3 Results

3.1 Two corpora

We have tested our method on two technical medium-size corpora. The first one, the Nuclear Technology Corpus (NTC) of EDF, is of about 52,000 words. The second one, the Coronary Medicine Corpus (CMC), is of about 60,000 words. It was built for the European MENELAS project (Zweigenbaum, 1994) and is used for pilot studies in terminology extraction[8].

3.2 A visual map of concepts and relationships

Even if no ontology can be fully automatically derived from a corpus (Habert and Nazarenko, 1996), the SYCLADE graphs can be used to bootstrap the building of the ontology of a domain.

The SYCLADE network gives a global view over the corpus which enables an alternate paradigmatic and syntagmatic exploration of the context of a word. The graph enables to identify the concepts, their possible typical properties, and also the relationships between the selected concepts.

The cliques bring out small paradigmatic sets of forms which, in a first step, can be interpreted as ontological classes reflecting concepts. The arc labels then help to refine those classes by adding some of the surrounding words which are not part of the clique but which nevertheless share the most significant or some similar contexts. From the clique {*stenose, lesion, obstruction, atteinte*} (cf. fig. 3), one can build the class of affections which are located in the body as {*plaque, occlusion, stenose, lesion, calcification, obstruction, atteinte*}. Similarly, from the graph of the CMC corpus, one can identify the classes of body sites {*artere, branche, reseau, ventricule, interventriculaire, carotide*}, of diseases {*maladie, artherosclerose*} and of chirurgical acts {*pontage, revascularisation, angioplastie*}.

Once these concepts are identified, their properties can be listed, by interpreting the labels of the links. The attribute of the localization of the affections is described through three kinds of modifiers (fig. 3): nouns (~ de {*artere, tronc*}), names of arteries (~ de {*carotide, interventriculaire*} and adjectives related to a specific artery (~ {*coronaire, coronarien, diagonal, circonflexe*}). The attribute degree of the affection is also revealed through {~ *significatif, non-significatif, severe, important, severite*}.

[6] The sub-graphs in which there is a path between every pair of distinct nodes.

[7] The sub-graphs in which there is a path between each node and *every other node* of the graph.

[8] Groupe Terminologie et Intelligence Artificielle, PRC-GDR Intelligence Artificielle, CNRS

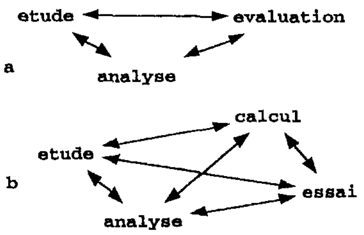

Figure 4: Polysemy of *etude*

Last, relationships between concepts can be extracted, such as the"part-of" relation between *tronc* and *artere*, and *segment* and *artere* (fig. 3).

3.3 Distinguishing word meanings

Polysemy and quasi-synonymy often makes the ontological reading of linguistic data difficult. However, through cliques and edge labels, the SYCLADE structured and documented map of the words helps to capture the word meaning level.

Among a set of connected words where w is similar to w_i and w_j, cliques bring out coherent subsets where w_i and w_j are also similar to each other. We argue that the various cliques in which a word appears represent different axes of similarity and help to identify the different senses of that word. For instance, in the whole set of words connected to *etude* (*study*) in a strongly connected component of the NTC graph (*analyse, evaluation, resultat, presentation, principe, calcul, travail...*), some subsets form cliques with *etude*. Two of those cliques (resp. *a* and *b* in fig. 4 – threshold of 7) bring out a concrete and a more theoretical use of *etude*.

The network also enables to distinguish the uses of quasi-synonyms such as *coronaire* and *coronarien* in the CMC corpus. Even if they are among the most similar adjectives (7 shared contexts) and if they belong to the same clique {*coronaire, coronarien, diagonal, circonflexe*}, the fact that *coronarien* alone is connected to evaluation adjectives (*severe, significatif* and *important*) shows that they cannot always substitute to each other.

4 Towards an adequate similarity estimatation for the building of ontologies

The comparison with the similarity score of (Hindle, 1990) shows that SYCLADE similarity indicator is specifically relevant for ontology bootstrap and tuning. Hindle uses the observed frequencies within a specific syntactic pattern (subject/verb, and verb/object) to derive a cooccurrence score which is an estimate of mutual information (Church and Hanks, 1990). We adapted this score to noun phrase patterns.[9] However the similarity measures based on cooccurrence scores and nominal phrase patterns are less relevant for an ontological analysis. The subgraph of the chirurgical acts words, which is easy to identify from the SYCLADE graph (fig. 5a), is split in different parts in the similarity graph (fig. 5b). This difference stems from the fact that this cooccurrence score overestimates rare events and underlines the collocations specific to each form.[10] For instance, it appears that the relationship between *stenose* and *lesion*, which was central in figure 3, with 8 shared contexts, almost diseappears if one considers the number of shared cooccurrences. Therefore, similarity measures based on cooccurrences and similarity estimation based on shared contexts must not be used in place of each other.

As opposed to Hindle's lists of similar words which are centered on pivot words whose neighbors are all on the same level, in SYCLADE graphs, a word is represented by its role in a whole syntactic and conceptual network. The graph enables to distinguish the various meanings of words, a crucial feature in the ontological perspective since the meaning level is closer to the concept level than the word level. In addition, the results are clear and more easily interpretable than those given by a statistical method, because the reader does not have to supply the explanation as to why and how the words are similar.

The building of an ontology, which is a time-consuming task and which cannot be achieved automatically, can nevertheless be guided. The SYCLADE graphs based on shared contexts can facilitate this process.

[9] For instance, for N_1PN_2

$$Cooc_{N_1PN_2} = \log 2\frac{\frac{f(N_1PN_2)}{k}}{\frac{f(N_1)}{k}.\frac{f(N_2)}{k}}$$

where $f(N_1PN_2)$ is the frequency of noun N_1 occurring with N_2 in a noun preposition pattern, $f(N_1)$ is the frequency of N_1 as head of any N_1PN_x sequence and $f(N_2)$ the frequency of N_2 in modifier/argument position of any N_xPN_2 sequence and k is the count of N_xPN_y elementary trees in the corpus. $Cooc_{NAdj}$ and $Cooc_{AdjN}$ are similarly defined.

[10] The various cooccurrence scores retrieve sets of collocations which are sharply different from the contexts shown by SYCLADE connected components. The collocations which get the greatest cooccurrence scores seem to characterize medecine phraseology (*facteur (de) risque, milieu hospitalier*) but not the coronary diseases as such.

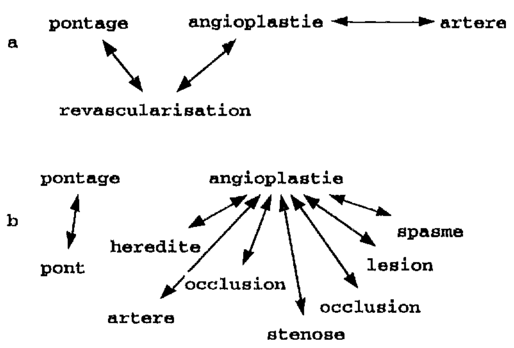

Figure 5: Similarity among the chirurgical act family

Acknowledgments

We thank Christian Jacquemin (IRIN), Didier Bourigault, Marie-Luce Herviou, Jean-David Sta (DER EDF), Marie-Hélène Candito (TALANA) and Sophie Aslanides (ELI) for their remarks on a previous version of this paper. We are very grateful to Serge Heiden (ELI), who has developed GraphX (`ftp://mycroft.ens-fcl.fr/pub/graphx/`), the graph interactive handling software that enabled us to visualize and handle the SYCLADE graphs.

References

Roberto Basili, Maria-Teresa Pazienza, and Paola Velardi. 1994. A "not-so-shallow" parser for collocational analysis. In *Proceedings of Coling'94*, pages 447–453.

Peter A. Bensch and Walter J. Savitch. 1995. An occurrence-based model of word categorization. *Annuals of Mathematics and Artificial Intelligence*, 14:1–16.

Didier Bourigault. 1993. An endogenous corpus-based method for structural noun phrase disambiguation. In *6th European Chapter of the Association for Computational Linguistics.*

Kenneth W. Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, march.

Béatrice Daille. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques.* PhD Thesis, Paris VII University, february. Supervisor: Laurence Danlos.

Gregory Grefenstette. 1994. *Exploration in Automatic Thesaurus Discovery.* Kluwer Academic Publishers.

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

Benoît Habert and Adeline Nazarenko. 1996. La syntaxe comme parche-pied de l'acquisition des connaissances. In *Actes des Journées d'Acquisition des Connaissances*, Sète, May.

Benoît Habert, Philippe Barbaud, Fernande Dupuis, and Christian Jacquemin. 1995. Simplifier des arbres d'analyse pour dégager les comportements syntaxico-sémantiques des formes d'un corpus. *Cahiers de Grammaire*, (20).

Donnald Hindle. 1990. Noun classification from predicate-argument structures. In *Proceedings of the Association for Computational Linguistics'*, pages 268–275.

Douglas P. Metzler and Stephanie W. Haas. 1989. The constituent object parser : Syntactic structure matching for information retrieval. In *Proceedings, 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'89)*, pages 117–126, Cambridge, MA.

Naomi Sager, Carol Friedman, and Margared S. Lyman (editors). 1987. *Medical Language Processing : Computer Management of Narrative Data.* Addison-Wesley.

Franck Smadja. 1993. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–177, march. Special Issue on Using Large Corpora: I.

Alan Smeaton and P. Sheridan. 1991. Using morpho-syntactic language analysis in phrase matching. In *Proceedings RIAO'91*, pages 415–429.

Pierre Zweigenbaum. 1994. MENELAS: an access system for medical records using natural language. *Computer Methods and Programs in Biomedicine*, 45:117–120.