Corpus-based Extension of a Terminological Semantic Lexicon

A. Nazarenko*, P. Zweigenbaum**, B. Habert† and J. Bouaud**

* Laboratoire d’Informatique de Paris-Nord
Université Paris 13
Avenue J.-B. Clément, 93430 Villetaneuse, France
nazarenko@lipn.univ-paris13.fr

** DIAM — Service d’Informatique Médicale/DSI/AP-HP
& Département de Biomathématiques, Université Paris 6
91, Boulevard de l’Hôpital, 75634 Paris 13 Cedex, France
{pz,jb}@biomath.jussieu.fr

† UMR 8503 — EcoleNormale Supérieure de Fontenay St Cloud
31, Avenue Lombart, 92260 Fontenay-aux-Roses, France
bh@ens-fcl.fr

This paper addresses the problem of extending and tuning a terminological semantic lexicon to new domains and corpora. We argue that by relying on both a sublanguage corpus and a core semantic lexicon, it is possible to give an adequate description of the words that occur in the corpus. Our tuning method explores the corpus and gathers words that are likely to have similar meanings on the basis of their dependency relationships in the corpus. The aim of the present work is to assess the potential for classifying words based on the semantic categories of “neighbors”. The tagging procedure is tested and parametrized on a rather small French corpus dealing with coronary diseases (85,000 word units). This method is systematically evaluated by creating and categorizing artificial unknown words. Although word semantic categorization cannot be fully automated, the results show that our tagging procedure is a valuable help to account for new words and new word usages in a sublanguage.

1. Introduction

In medical information processing, medical nomenclatures and thesauri (Cimino 1996) are being used extensively, from classifying patient data for statistical purposes to decision-support (Mussen & van Bemmel 1997). However, the constant changes in techniques and approaches prevent these fundamental resources from ever being complete. Moreover, variations in terminology can be observed for a given specialty in different places. There is therefore a constant endeavor to tune vocabularies and terminologies to account for new words and new word usages.

Controlled terminologies help to record and access medical information. Semantic lexica, which associate (syntactic and) semantic categories to words, are necessary as well. They help to match controlled terms to the expressions that actually occur in corpora, for instance via synonymy relationships between heads or modifiers, or via semantic pattern-matching (McCray et al. 1994; Naulleau 1998). This contributes to terminology construction and update, by allowing the acquisition of terms from corpora and their
linkage with existing controlled terms (Nelson et al. 1998). This also facilitates terminology usage, for instance in document indexing or in a controlled language setting, since it helps identify the suitable controlled terms for an initial expression (Sneiderman et al. 1996).

Here we address the problem of extending a semantic lexicon for a given sublanguage. We use an existing terminology as a “seed”, as it offers categories which correspond more tightly to the salient notions and relations in the domain than a general semantic lexicon such as WordNet (Fellbaum 1998). We argue that by relying on both a corpus and a core semantic lexicon based on a terminology, it is possible to tune, to extend, and to adjust this lexicon to give an adequate description of the words that occur in the corpus. We rely on a method which explores a sublanguage corpus and gathers words that are likely to have similar meanings. We derive a set of a priori semantic classes from a terminology for this sublanguage. We then propose a method to assign a semantically unknown word to the class that contains the words that are most similar to it on the basis of their observed behavior in the corpus.

Till now, we have tested on single words instead of multi-word terms, and with a small set of domain-specific, high-level classes rather than with a multi-level hierarchy. These constraints allow us to test the method more easily. But we claim as well that such coarse-grained categories are more relevant to semantic retrieval than finer ones. The aim of the present work is to assess the potential for classifying words based on the semantic categories of “neighbors” identified through normalized syntactic distributional properties: their dependency relationships in the corpus.

The remainder of this section presents our experimental setting. Section 2 describes our general approach to lexicon tuning: we rely both on corpora and lexical resources. Section 2 presents our tuning method which complements and adjusts an existing lexicon with respect to a corpus. It is based on the propagation of semantic tags in ZELLIG graphs. The first results called for a systematic evaluation, which is described in section 0 and which helped to parametrize our method. Results are reported in section 5. The last sections discuss the results (5) and the potential applications (7) of this work.

### 1.1. The MENELAS corpus

These experiments were carried out on a French corpus which was initially designed for the European project ZELLIG (Zweigenbaum 1994), devised for the development of a system for analyzing medical reports. The ZELLIG corpus gathers texts dealing with coronary diseases (patient discharge summaries, discharge letters and a handbook on coronary angiography). It is a rather small corpus (85,000 word units) as compared to the very large corpora which are sometimes required to test NLP techniques. However, this size is typical of the available technical corpora, which are often too small for statistical methods and nevertheless too large for manual analysis.

### 1.2. The SNOMED terminology

The experiments also made use of an existing terminological resource: the “Systematized Nomenclature of Human and Veterinary Medicine”, version 3, also called “SNOMED International” (Côté et al. 1993). SNOMED is a multi-axial, hierarchical terminology with 11 high-level “axes”: Topography (T), Morphology (M), Function (F), Living Organisms (L), Chemicals, Drugs, and Biological Products (C), Physical Agents, Activities, and Forces (A), Occupations (J), Social Context (S), Diseases/Diagnoses (D), Procedures (P), and General Linkages/Modifiers (G). There is a French translation for a subset of this terminology, the Microglossary for Pathology (12,500 terms) (Côté 1996). SNOMED is one
Corpus-based Extension of a Terminological Semantic Lexicon

of the medical terminologies with the best coverage of clinical terms (Chute et al. 1996). We have chosen it as a compromise between clinical coverage and availability in French. We limited the richness of the SNOMED hierarchy to its 11 high-level categories without considering their subcategories. This choice was a compromise between the size of the tagset and the reliability of the tagging, to which we return below (see section 2.3).

As test material for our method, we created a semantic lexicon where each relevant word in the ZELIG corpus is assigned one or more of these categories. This lexicon totals 1,902 lemmas. It can be seen as a simplified version of a subset of the terminology relevant for the corpus, with only single-word terms.

1.3. Tuning the SNOMED lexicon to the MENELAS corpus

In previous work (Habert et al. 1996), we explained how ZELIG discovers similarities between words according to the contexts they share within a given domain. The present work attempts to show that this corpus-based knowledge (similarities) and a semantic lexicon (obtained from SNOMED) can be combined to tune this lexicon to the domain of coronary diseases as described in the ZELIG corpus. The similarities between words help to tag words that are unknown to the semantic lexicon and to detect ill-tagged words. The tagset consists of the high-level SNOMED axes used as general semantic categories. Through systematic experiments, the present work attempts at quantifying the extent to which this process succeeds in proposing a correct category for a given word of the corpus as several parameters of the method are submitted to variation.

2. A corpus- and knowledge-based approach

2.1. The need for semantic knowledge

Many attempts have been made to infer semantic categories from corpora using purely endogeneous methods, i.e., using no semantic knowledge apart from the one driven from the corpus itself. Using the syntactic distribution of words in a corpus, as obtained by a parser, to group words into semantic classes, has already been the subject of past research (Hirschman et al. 1975). The increased availability of large corpora as well as of robust NLP techniques and tools has revived this type of work. As noted in (Grefenstette 1994), one can usually distinguish three steps in discovering semantic affinities between words:

- First-order techniques examine the local context of a word attempting to discover what can co-occur with that word within that context. Second-order techniques derive a context for each term and compare these contexts to discover similar words or terms.
- Third-order techniques compare lists of similar words or terms and group them along semantic axes.

After cooccurrence and similarity relations, third-order techniques aim at bringing out equivalence distributions between words, that is to say, at building semantic classes. However, no coherent and interpretable semantic classes can be built on purely endogeneous grounds. Some of the classes obtained by Bensch & Savitch (1995) or McMahon & Smith (1996) seem to be semantically sound at first sight but one must adjust their boundaries and check their consistency to turn them into actual categories. A similar conclusion can be drawn for the similarity subgraphs of (Habert et al. 1996).

Some external source of knowledge is therefore required to name semantic classes, either human interpretation (as in Faure & Nedellec 1998) which presents the construction of word classes as an interactive process) or a preexisting lexical resource.
2.2. Specialized vs. general-purpose lexical resources

General lexical resources such as ordinary language dictionaries or thesauri are often used. Many works have exploited WordNet, Roget’s Thesaurus or traditional, albeit machine-readable, dictionaries for word sense disambiguation on newspaper corpora for instance. These lexical resources have also been exploited for more technical corpora. In sublanguages, words often have very specific and unusual meanings; polysemy does exist but it is often limited. (Basili et al. 1997; Basili et al. 1998) propose a corpus-based method to adapt WordNet or LDOCE to a sublanguage: all non-attested word senses are eliminated; on the other hand, if there is corpus evidence of an undescribed word use, a corresponding new sense is added. Hamon et al. (1998) exploit the French dictionary Le Robert to identify synonymous pairs in the list of terms of a document. Although it is often assumed that general lexical resources are of no use for the processing of technical texts, these first experiments have shown that this approach is relevant as soon as lexical information is controlled by corpus attestations.

As far as technical corpora are concerned, one usually looks for specialized lexical sources, if there are some (Morin 1998; Hamon et al. 1999). However, adequate lexical knowledge sources are seldom available. From one experiment to another, the domains are always slightly different. Charlet et al. (1996) underline that knowledge bases depend on the task they have been designed for. This is equally true for lexical knowledge bases. Moreover, in a technical domain such as medicine where terminological uses change with place and time, terminologies are quickly out-of-date.

Our approach is close to the works based on specialized lexical sources as we make use of similar lexical information. However, our aim is closer to Basili’s. We argue that specialized resources, as well as more general ones, need to be updated for each specific application (domain and task). Even if our tuning method differs from that of Basili et al. (1997), our approach is similar to theirs.

2.3. Few general semantic categories

For our tuning procedure, we deliberately took a small set of general semantic tags as in (Basili et al. 1993). As explained in the introduction, we worked with the 11 high-level categories (Topography, Morphology, Function, Living Organisms…) of the SNOMED terminology. Basili et al. (1993) argue that a set of a dozen tags is a good starting point to perform manual word sense disambiguation. A refined tagset would increase the contrast between expert judgements. Choosing a tag is also less costly if the tagset is smaller. In our case, the tagging procedure relies on a vote among the categories of similar words. Increasing the tagset would spread out semantic information for a given word. (Agirre & Rigau 1996) came to a similar conclusion while exploiting WordNet for word sense disambiguation: instead of relying on the small, numerous and always questionable synsets, they rather take the “general categories” represented by WordNet files. More recent work by Basili et al. (1997) also relies on these general WordNet categories. As will be shown below, starting with general categories still gives clues for further subcategorization.

3. A tuning method based on word dependency similarities

Our reference corpus is first processed with the ZELLIG suite of tools (Habert et al. 1996) which builds a similarity graph of the corpus words (actually, lemmas). The graph nodes are the words: two words are connected if they are considered similar. These words are
tagged according to the top SNOMED categories. The graph is then used to tag unknown words or to detect erroneous tags.

3.1. Building a similarity graph of the corpus

ZELLIG follows the above mentioned three-step process to discover similarities between words within a given domain according to the contexts they share (Nazarenko et al. 1997). It relies on normalized syntactic noun phrases (NPs) as local contexts for the first-order step. It uses parse trees retrieved by noun phrase extractors (in the present experiment, LEXTER (Bourigault 1993)). It leaves as such the attachments produced by the parser. ZELLIG automatically reduces the numerous and complex noun phrases provided by LEXTER to elementary dependency trees, which more readily exhibit the fundamental binary relations between content words (Harris 1991).

A first step consists in rewriting these parse trees so as to get normalized binary trees: each level associates a head with either an argument or an adjunct. This normalization process facilitates the extraction of dependency relationships. The second step consists in extracting elementary dependency trees, in which a head (in this case a noun) is modified by another content word (a noun, possibly with a preposition or an adjective). For instance, from the parse tree for “signe périphérique de décompensation cardiaque” (figure 1), ZELLIG yields the set of elementary trees of figure 2. Note that this normalization process simplifies the contexts, reduces their diversity and increases their number: from one complex parsed context for décompensation, the normalization produces two elementary contexts (trees b and c).

---

**Figure 1:** Complete parse tree for “signe périphérique de décompensation cardiaque”.

(a) “signe de décompensation”

(b) “signe périphérique”

(c) “décompensation cardiaque”
Figure 2: *Elementary dependency relations extracted from the parse tree for "signe périphérique de décompensation cardiaque".*

Second-order affinities show which words share the same contexts. For instance, the following words can replace *signe* in tree *b:* récidive (1 occurrence), épisode (2 occurrences), etc. All these words can occur in the same context: a NPNP tree, whose second noun is décompensation and whose preposition is de.

A graph is computed by ZELLIG to exhibit salient similarities. The words constitute the nodes. An edge corresponds to a certain amount of shared contexts, according to a given measure and a chosen threshold, which vary in the following experiments. Figure 3 shows the immediate edges around the word *artère* and brings out the words that are similar to *artère* at threshold 9. For this experiment, the similarity measure is CTXT (see section 4.2). Each edge is labelled with the number of context types shared by the two words it links. In this example, the words which share less than 9 contexts with *artère* are not considered similar to it and do not appear. On figure 3, SNOMED categories (/T, /M) have been added (projected) to nodes, as explained below (section 0).
As a third-order technique, ZELLIG first computes the strongly connected components (the sub-graphs in which there is a path between every pair of distinct nodes) and the k-cliques (the sub-graphs in which there is an edge between each node and every other node of the graph). These are the most relevant parts of the graph on topological grounds. The underlying intuition is that a connected component relates neighboring words with semantic intersections and that the cliques tend to isolate sets of synonyms, of antonyms or scalar groups of words. Examples of a connected component and a clique are presented on figures 4 and 5. Figure 5 shows on each edge the contexts shared by the two nodes that it links. For the sake of clarity, these contexts have been removed from the edges in figure 4.

The connected components are further used in the tagging method described below.
3.2. **Tuning a lexicon**

The ZELIG graph represents a map of the corpus words which can be confronted to an existing categorization in order to complete, correct, specialize and update it. For instance, as explained in the introduction, the SNOMED Microglossary (Côté 1996) only covers a part of the corpus lemmas. Our tagging method therefore starts with a categorization of only a part of these lemmas. The SNOMED categories of known words are projected on the ZELIG graph nodes. The unknown words are initially untagged. We chose the following tagging heuristic: *given an untagged lemma in a connected component, its semantic category is chosen by majority (according to a given voting procedure, see section 4.4) of those of its neighbors*. The categories are thus propagated through the edges from the already tagged lemmas (known words) to the untagged ones (unknown words). As a trivial case, untagged lemmas in a homogeneous connected component get the semantic category of the rest of the component. In contrast, some lemmas may remain untagged if there is no majority among their neighboring tags. The same propagation process also brings out tagging inconsistencies if a word has a tag which contradicts those of its neighbors.

3.3. **First experiment**

In a first experiment (Nazarenko et al. 1997), we only categorized 937 lemmas (those that were present in the original MENELAS semantic lexicon). We then projected these categories on the lemmas at the nodes of the connected components and of the cliques obtained at thresholds 5 and 10. On the connected component of figure 6, many lemmas can be tagged according to the SNOMED nomenclature and a few ones (e.g., apical or artériel) are unknown. Giving unknown words the category which has the majority among their neighbors assigns category G to /apical, /postéromédiolatéral, /distale and /récent (unanimously for the first 3, 2 against 1 for /récent), and this choice is correct according to our knowledge of the domain; /artériel obtains a tie with 1 against 1 (G/gauche T/coronarien), and therefore does not get tagged. Considering a set of 143 lemmas (corresponding to all the connected
components but the largest) and starting with 87 already tagged lemmas, this heuristic tagged 46 of them and left untagged 10 lemmas unknown to SNOMED. 38/46 taggings were consistent with our knowledge of the domain, 4 were erroneous, and 4 raised a doubt which required to go back to the corpus.

4. Evaluating the tuning method through systematic experiments
To assess the performance of the method, we carried out systematic experiments and we evaluated the tagging results as several parameters of the method were submitted to variation.

4.1. Creating artificial unknown words
In the preceding experiment, it was difficult to give a proper evaluation of the method for two reasons. The first one is that words unknown to SNOMED in the ZELLIG graphs computed at the preceding thresholds for the MENELAS corpus are not very numerous (the problem of sample size). The second one is that it is sometimes difficult to determine which would be the proper tag for a given word (the problem of reference) since there is no gold standard for the problem.

Following a traditional evaluation procedure (cf. for instance (Grefenstette 1994) for the creation of artificial synonyms), we artificially created unknown words, i.e., words known to SNOMED but for which we removed the SNOMED tags, and tried to have their tags automatically guessed. We also varied the threshold, so that more words were involved. We consequently extended our test semantic lexicon to cover these lemmas. This resulted in an experimental sample of 1,902 lemmas, whereas the first one only contained 57 (actually) unknown words. We consider that a tag guessed by our method is correct if it corresponds to the one in our reference lexicon.

We then had a procedure examine each word in turn, assuming its category was unknown, guessing it from its neighbors, and comparing the guess with the reference category found in the lexicon. This evaluation protocol helps to test the method and tune its various parameters.

4.2. Computing a similarity value
In our first experiments with the ZELLIG suite of tools, we considered that two words were similar if the number of the different contexts they shared was above a given threshold. A context in which a word occurs more often may be considered stronger than one with
only a few occurrences, so that two words that often occur in a given common context
may be deemed more similar than two words that only occur once each in that context.
This is what the CTXO measure takes into account. Besides, words that occur with many
non-shared contexts may be considered less similar than words that only occur in shared
context. The JACCARD measure below can be used to model this behavior.6
Each word \( W_i \) occurs in a set of different contexts \( \{ctx_i^k\} \), each with a number
of occurrences \( |ctx_i^k| \). Given a word \( W_i \) and a potential neighbor \( W_j \), we compute a similarity
\( sim-X_{ij} \) between \( W_i \) and \( W_j \). We used as \( sim-X_{ij} \):

- their number of shared context types:
  \[ sim-CTX_{ij} = |\{ctx_i^k\} \cap \{ctx_j^k\}| \]
  (this is the initial similarity measure we used (Habert et al. 1996));
- the frequency of their shared context types:
  \[ sim-CTXO_{ij} = \sum_{k \in \{ctx_i^k\} \cap \{ctx_j^k\}} \min(|ctx_i^k|, |ctx_j^k|); \]
- a JACCARD measure (Saporta 1990) equivalent to the Dice coefficient (Dice 1945)
frequently used in information retrieval (Salton & McGill 1983; Frakes & Baeza-
Yates 1992) over context types:
  \[ sim-JACCARD_{ij} = \frac{|\{ctx_i^k\} \cup \{ctx_j^k\}|}{|\{ctx_i^k\} \cup \{ctx_j^k\}|}. \]

JACCARD is the only of the three to be a normalized similarity index. It normalizes
the number of shared context types by the total number of context types of
\( W_i \) and \( W_j \), so that words which are used in exactly the same contexts are favored over
words which occur not only in many shared contexts (sim-CTX), but also in many
distinct contexts.

Polysemous words are multiply tagged on graph nodes. As a crude method to take all
these tags into account, a polysemous word votes once for each of its categories.
Table 1 shows the similarity values for “décompensation”, “palpitation” and
“insuffisance”.

<table>
<thead>
<tr>
<th>arête</th>
<th>CTXT</th>
<th>CTXO</th>
<th>JACCARD</th>
</tr>
</thead>
<tbody>
<tr>
<td>décompensation, palpitation</td>
<td>2</td>
<td>2</td>
<td>0.25</td>
</tr>
<tr>
<td>décompensation, insuffisance</td>
<td>2</td>
<td>6</td>
<td>0.04</td>
</tr>
<tr>
<td>palpitation, insuffisance</td>
<td>3</td>
<td>3</td>
<td>0.05</td>
</tr>
</tbody>
</table>

4.3. **Pruning the graph**
Given a similarity measure, we then prune each couple of words whose similarity falls
below a given threshold.
Depending on the chosen threshold, two words may appear similar or not. The pruning
modifies the topology of ZELLIG graphs: when the threshold is raised, the number of
edges decreases, leading to smaller connected components. Figure 7 shows how raising
the threshold and changing the similarity measure affect the subgraph of
“décompensation”, “palpitation” and “insuffisance” and thus the whole tuning method.
Figure 7: Alterations to the subgraph of “décompensation”, “palpitation” and “insuffisance” due to similarity measure variations (for similar thresholds).

As there was no a priori reason for choosing one pruning threshold or another, we examined a relevant series of thresholds for each kind of measure: {1 2 … 19} for CTXT, {1 2 3 4 5 10 15 20} for CTXO, and {.00 .02 .04 .06 .10 .20 .30 .40 .50 .60} for JACCARD.

4.4. Ranking the categories of the immediate neighbors

Given a word \( W_i \) in the pruned graph, we have each of its immediate neighbors vote for its category, as found in the lexicon. We tested two vote aggregation methods:

− each vote is counted as one (unweighted);
− the vote of each neighbor \( W_j \) is weighted by its similarity \( \text{sim-}X_{ij} \) with word \( W_i \).

The first voting procedure takes the graph as a boolean similarity matrix whereas the second one takes similarity distances into account.

Votes for the same category are summed, and categories are ranked according to their cumulated votes. The top-ranked category is assumed to be the category of the word \( W_i \). For instance, on figure 3, two categories compete for the tagging of “artère”. Taking \( W_i = \text{artère} \), in the weighted aggregation scheme, gives the following two scores:
\[ \sum_{cat=T} \text{sim-CTX}_{ij} = 45, \]
\[ \sum_{cat=M} \text{sim-CTX}_{ij} = 23, \]
so that category T (Topography) is ranked first and M (Morphology) is ranked second. T is actually the correct SNOMED tag for "artère".

5. Results

Table 2 shows how many units were computed at each step of the method. LEXTER extracted 6,896 NPs in which 10,832 elementary trees were identified. Elementary trees mostly cover noun dependency constructs, so that some words (342, i.e., 15%) were left out at this stage. The resulting 21,664 contexts include 2,073 nouns, adjectives, proper nouns and past participles, plus 11 prepositions (0.5%, used as links in the contexts). Focusing on the 40,427 triples \{shared context, word_1, word_2\} again eliminates 204 lemmas (10%) that only occur in idiosyncratic contexts. In total, 23% of the NP lemmas did not reach the graph.

Once on the graph, lemmas were tagged according to the lexicon. This lexicon happened to lack 149 of the lemmas, so that the corresponding nodes were removed from the graph, which caused a few more nodes to become disconnected from the rest of the graph. 158 lemmas (8.5%) were lost at that stage, which resulted in a total diminution of 715 lemmas (29.5%) starting from the initial NP contents. The maximal recall the tagging step may obtain in this setting is thus 70.5%.

Table 2: Numbers of units at different processing stages (T = threshold).

<table>
<thead>
<tr>
<th>Unit</th>
<th>Units</th>
<th>Lemmas</th>
<th>Units</th>
<th>Lemmas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun phrase</td>
<td>6,896</td>
<td>4,566</td>
<td>2,402</td>
<td>2,426</td>
</tr>
<tr>
<td>Elementary tree</td>
<td>10,832</td>
<td>4,925</td>
<td>2,501</td>
<td>2,084</td>
</tr>
<tr>
<td>Context</td>
<td>21,664</td>
<td>3,590</td>
<td>2,166</td>
<td>2,073</td>
</tr>
<tr>
<td>Shared context</td>
<td>40,427</td>
<td>1,513</td>
<td>1,869</td>
<td></td>
</tr>
</tbody>
</table>

(a) Number of noun phrases and contexts  
(b) Number of nodes (lemmas in graph)

Given a word, our procedure ranks the categories of the neighbors, from most to least salient. The correct category for the word may get ranked first (rank 1), which is the desirable situation. It may also be ranked second, third, etc., or may not be present at all in the neighbors. There may also be a tie between the first and following categories.
We examine the distribution of ranks among the words of the graph obtained at the various thresholds for each measure (\textit{CTXT}, \textit{CTXO}, \textit{JACCARD}) and with each ranking method (unweighted, weighted), both as a percentage of the total number of words in the graph (relative) and as an absolute number of words. The percentage of rank 1, which is the ratio of the number of correctly categorized words (rank 1) over the total number of words considered at that threshold, corresponds to the categorization precision. Recall is computed as the absolute number of rank 1 words over the total number of words in the corpus noun phrases (i.e., 2,426 words). Table 3 provides precision and recall figures for the six combinations of similarity and weighting choices and for the main thresholds.

6. Discussion

6.1. Variation of precision and recall with parameter settings

\textbf{Weighting} increases clearly (by 5 to 10 \%) both precision and recall for all methods and at most thresholds. The semantic category conductivity of \textit{Zellig} graph edges – that is, they allow the possible propagation of a semantic category to a word from its neighbors – is thus proportional to their similarity strengths.

\textbf{Precision} varies with the \textbf{threshold} for \textit{CTXT}: it rises from 46/51 \% (depending on whether weighting is on/off) at threshold 1 to an optimum of 70/75 \% at thresholds 8–9, and then decreases. \textit{CTXO} displays slightly less variation, while \textit{JACCARD} is remarkably stable around 44–47/49–52 \% at thresholds 0.00–0.20 and keeps a 43 \% precision at the maximum threshold.

\textbf{Recall}, obviously, quickly decreases with the \textbf{threshold} for \textit{CTXT} and slightly less for \textit{CTXO}, while the decrease for \textit{JACCARD} is much slower.

In summary, weighting seems to be beneficial in all cases, as in any of the studies of similarity and clustering performed over the past 50 years. Maximum recall requires to use a low threshold for all methods (except for \textit{JACCARD}, but using a higher threshold does not increase precision in that case). At low thresholds, the three methods do not display really different precision scores.

We are thus facing the traditional trade-off between precision and recall: the higher the threshold the more reliable the assigned categories. It could improve the method to assign categories to graphs of decreasing thresholds and to assign reliability measures to these assigned categories.
It is interesting to observe that, with CTXT, which displays rather better precision with some non-minimal thresholds, the assigned categories are fairly stable when the threshold is decreased. 93.5% of the correct categories found at threshold 9 (unweighted) are also found at threshold 8. 100% and 77.7% of the correct categories found at thresholds 6 and 3 respectively can also be found with a lower threshold. Only 10% of the words that are correctly categorized with a threshold higher than 2 are not correctly categorized with the graph of threshold 2. This leads us to favor recall over precision.

On the whole, among the 2,426 words in the SNOMED corpus reduced to NPs, 1,711 (70.5%) appear in the threshold 1 CTXT graph, the missing words being words with such a specific use that they share no context with any other word. At threshold 2, 672 (27.6%) words remain. Among these 672 words, 49.6% automatically receive a correct category through the unweighted scheme. If one considers that only 602 words can actually be categorized (70 should not be categorized because of a tie), we end up with a 55.3% precision. This global percentage is encouraging if one considers that randomly choosing a category among a set of 11 categories would yield a score of 9%. However, assigning G to all words gives a better baseline precision of 301/672 or 44.8%.

6.2. Variation of precision and recall with SNOMED categories

In fact, the quality of the categorization procedure varies with the SNOMED category, ranging from 14.3% for L (Living Organisms) to 65.4% for G (General Linkages/Modifiers).

Table 4 shows that the categorization precision is roughly correlated with the number of words in each category, the largest categories G (General Linkages/Modifiers) and F (Function) obtaining the highest precision. The same observation was made in other experiments in which we worked with categorizations other than SNOMED (El Bouchikhi 1999). This is a known property of memory-based reasoning systems (Berry & Gordon 1997).

Table 4: Results per SNOMED category (CTXT, unweighted, threshold 2).

<table>
<thead>
<tr>
<th>Category</th>
<th># of words</th>
<th>Correctly categorized</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Local continuity tagging</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nb</td>
</tr>
<tr>
<td>Physical Agents, Activities, and Forces (A)</td>
<td>12</td>
<td>2</td>
</tr>
<tr>
<td>Chemicals, Drugs, and Biological Products (C)</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>Diseases/Diagnoses (D)</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>Function (F)</td>
<td>110</td>
<td>61</td>
</tr>
<tr>
<td>General Linkages/Modifiers (G)</td>
<td>301</td>
<td>197</td>
</tr>
<tr>
<td>Living Organisms (L)</td>
<td>7</td>
<td>1</td>
</tr>
<tr>
<td>Morphology (M)</td>
<td>58</td>
<td>10</td>
</tr>
<tr>
<td>Procedures (P)</td>
<td>63</td>
<td>22</td>
</tr>
<tr>
<td>Social Context (S)</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Topography (T)</td>
<td>87</td>
<td>33</td>
</tr>
<tr>
<td>Occupations (J)</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Overall</td>
<td>672</td>
<td>333</td>
</tr>
</tbody>
</table>

Assigning a probability to each category according to its frequency gives a better baseline than random choice for comparison. Table 4 shows that, for all categories, the actual categorization precision is clearly higher than the probability score.

On the contrary, the categorization results are not correlated with the density of the graph. A word can be correctly or incorrectly categorized independently of the number of
neighbors it has in the graph. Therefore nothing prevents the categorization of words that have specific uses and share contexts with a single word, provided there is at least one shared context.

Our categorization procedure could also help humans to structure the larger categories into smaller ones. The vote brings out the major category, i.e., the most salient category among neighbors, but may also bring out a minor category, the second most salient one. For instance, among the G-categorized words (General Linkages/Modifiers), one can contrast the $G_T$ (G + toponymy) and $G_P$ (G + procedure) subgroups:

1. $G_T$ antérieur antéro-apical apical collateral minime significatif sévère ...
2. $G_P$ actuel année immediat jour mois possibilité réalisation réapparition ...

which may prove relevant for further subcategorization.

6.3. Limitations
The analysis of erroneous categories brings out the major cases of categorization ambiguity (apart from the words which can belong to several categories – we do not focus on these polysemy phenomena). The general category G is responsible for a large number of errors, especially for the modifiers: adjectives ischémique and systolique and nouns such as mouvement, which can be viewed as a “support noun”, all get categorized as G whereas they generally correspond to SNOMED functions (F). In these cases, purely syntactic similarities seem to be stronger than semantic characteristics. However, the much larger representation of the G category in the lexicon also increases the probability for each lemma to have G neighbors (Berry 1997). Note that Hirschman et al. (1975) carefully set apart words with specific syntactic behaviors so that their clustering method could work properly.

We mention three limitations in these experiments. First, we used semantic categories originally suited for terms in the SNOMED nomenclature. When we built our single-word test lexicon with these categories, it was easier to manually categorize words that occur as term heads in the SNOMED Microglossary than those that only occur as modifiers: many of the latter were therefore categorized in the G category (general linkages and modifiers). As a result, term heads are generally better categorized than the other components, whose semantic roles heavily depend on the noun they modify. This mainly affects precision. Secondly, the noun phrase extractor whose output was normalized by ZELLIG, in this experiment LEXTER, filters its results: it only yields phrases which can function as terms. The syntactic pattern and the semantic classes of the components must be compatible with such a role. For instance, sténose serrée à la fin du tronc commun is filtered out, as the complex preposition à la fin du does not occur in terms. This only affects recall. Last, the morpho-syntactic tagger used by LEXTER is not error-free. Tagging errors imply erroneous attachments and phrases, which affect both precision and recall.

7. Perspectives

7.1. Further experimentation
We noted above that whereas weighting was useful, using a threshold was not really desirable, and that the different similarity measures tested do not bring drastic changes at low thresholds.

In the reported experiments, the graphs are first built, then pruned and finally categorized according to the SNOMED axes. However, SNOMED categories could be incorporated at earlier steps, into either the graph pruning or the graph building processes. In the first
case, once the graph is built, the contexts are categorized. Similarity could therefore be no longer based on context types or occurrences but on categorized context types or occurrences. In the second case, context categorization would be done before the graph is built and would thus modify the graph topology itself by highly increasing the edge number. Categorizing the contexts is a generalization process (Grishman & Sterling 1994) that decreases the contextual diversity and modifies the resulting similarities. Incorporating semantic categories at earlier steps would further normalize the dependency contexts.

7.2. Terminology tuning and updating
This categorization method, originally developed for tuning an incomplete nomenclature for a given technical corpus, can have various other applications. It could be used to progressively enrich a nomenclature from incoming texts, i.e. to incorporate the texts produced by one or several hospitals or departments on a monthly, weekly or daily basis. Actually, our procedure can even categorize some hapaxes: it can work at low thresholds since the categorization of a word does not require it to have a large number of neighbors. Even if only few unknown words appear in each group of texts, we argue that an automatic categorization process is necessary. Manual categorization is not only costly, it is also not fully reliable. In a technical domain where terminology is changing according to place and/or time, it may be difficult to manually identify the category of an unknown word which could be a deceptive cognate or to detect the new uses of an already known word.

A different kind of application would consist in enriching the nomenclature itself. Categorizing unknown words extends its coverage and we have seen how the voting results can help to subcategorize a general category such as the SNOMED G axis. However, such an application requires that our method be tested on larger corpora.

7.3. Semantic categorization
This tagging procedure produces a semantic categorization of the words in the corpus. Such a categorization is useful both for terminological applications (control and structuration of terminologies, term definition, etc.) and for NLP ones (syntactic disambiguation, acquisition of selectional patterns, acquisition of extraction patterns, etc.). This process is designed to aid human terminological work. It cannot be fully automated but we have shown that, starting with few general semantic categories and a core lexicon which are available for many domains or which can be built for a specific purpose (Basili et al. 1993), the tagging procedure can help extend and refine the initial lexicon.

8. Conclusion
We presented a method for tuning a terminological semantic lexicon to new domains and corpora. The reported experiment aims at tuning a semantic lexicon derived from the SNOMED medical nomenclature to the ZELLIG corpus. The method is based on the ZELLIG suite of tools and exploits local continuity in graphs of dependency-based similarities. A tagging procedure is designed that propagates semantic tags through the graph edges from known words to unknown ones. Systematic experiments of categorizing artificial unknown words have been carried out. They help evaluate the whole tuning method. In the context of our experiment, the tagging procedure consists in choosing one tag among 11. The results are fairly good (almost 50% precision) compared with a baseline tagging choosing the most probable tag.
but they do not allow an automatic tagging. Syntactic dependencies are interesting clues for semantic categorization but tagging inconsistencies call for human analysis. For instance, if a semantic constraint is violated, i.e. if the tag of a word is inconsistent with those of its neighbors, one cannot guess from the corpus alone whether this is due to an initially erroneous tag or to an ellipsis. Our method is therefore designed as an aid for terminological building and updating. Result comparison has shown that thresholding the similarity graphs is useless and that a normalized similarity measure does no better than more trivial ones, at least on a small corpus. ZELLI graphs are thus used as a boolean similarity matrix where two lemmas are connected if they share at least one context. Weighting only benefits the voting procedure: it gives better results for the propagation of semantic categories. However, we have shown that the voting can be performed even for words which have only few neighbors. These results have important consequences for the terminological applications of our method as it can be applied to small corpora and help to categorize unfrequent words.

Acknowledgements
We thank Dr. R.A. Côté for graciously providing us a copy of the French version of the SNOMED Microglossary for Pathology.

References


Corpus-based Extension of a Terminological Semantic Lexicon 19


---

For instance, the refinements of *WordNet* synsets lead to difficulties when there are used in query extension (Voorhees 1998).
A lemma subsumes the morphological variants of a word. For instance, the masculine singular form is chosen for an adjective: *marginal* is the lemma for *marginale*, *marginaux*, and *marginales*.

Some of these single-word terms existed as is in the Microglossary; other words occurred in multi-word terms of *SNOMED*; and some did not occur at all in the *SNOMED* Microglossary. The latter two kinds of words were categorized manually to build this test semantic lexicon.

LEXTER uses non ambiguous contexts in order to choose between competing attachments. For instance, in the parse tree for “*signe périphérique de décompensation cardiaque*” (figure 1), “*cardiaque*” is attached to “*décompensation*” and not to “*signe*” because “*décompensation cardiaque*” is found in the corpus and “*signe cardiaque*” is not.

Two techniques were used successively. The first technique consists in “undoing” the rules which have been applied to derive the parse tree and to check after each simplification phase whether the current result exhibits an elementary dependency between content words. The second technique uses underspecified tree descriptions as an approximate tree matching device (Vijay-Shanker 1992).

All the context types do not necessarily have the same significance. Let us consider for example two contexts of *artère*, type d’*artère* (a) and *diamètre de l’artère* (b). The second one is far more semantically informative than the first one which is a kind of complex determiner. b is therefore more discriminant for word similarity measures. These linguistic properties may vary from one corpus to another. They are difficult to formalize and to take into account as such. They are regularly approximated quantitatively, however, for example in information retrieval systems.

Category S (Social Context), for which no correct category was found, does not have a significant size; the 12 J words (Occupations) in our corpus happened to all be hapaxes – that is words occurring only once – with specific contexts.