Extending an Existing Specialized Semantic Lexicon

B. Habert, A. Nazarenko, P. Zweigenbaum, J. Bouaud
LIMSI & UMR 9952 — École Normale Supérieure de Fontenay St Cloud – bh@ens-fcl.fr
LIPN — Université Paris 13 – nazarenk@lipn.univ-paris13.fr
DIAM — SIM/AP-HP & Dépt. Biomathématiques U. Paris 6 – {pz, jb}@biomath.jussieu.fr

Abstract

There is a constant need to extend and tune specialized vocabularies to account for new words and new word usages. This paper addresses the issue of characterizing the semantic class of such words. We test the hypothesis that the analysis of word distribution in a representative corpus, as obtained by robust NLP tools, can help identify words with similar meanings, and to decide on the most likely category for a given word based on the categories of its neighbors. We report on an experiment with a moderate-size corpus of patient discharge summaries collected during the MENELAS project, taking as categories the high-level axes of the SNOMED nomenclature, and processing the corpus with the ZELLIG suite of tools. We attempt to quantify the extent to which this process succeeds in proposing a correct category for a given word of the corpus while we vary several parameters of the method. The percentage of correctly categorized words (*precision*) ranges between 50 and 75 %, while the best percentage of categorized words (*recall*) is 37 % for the whole categorization process. Categorization results are significantly above chance, but not sufficient for a fully-automated process. We discuss possible uses of such a categorization help and identify further directions for improvement.

1 Introduction

As many technical words or word meanings cannot be found in general semantic databases such as a dictionary or WordNet, Natural Language Processing (NLP) in specific domains requires specialized semantic lexica. In a domain such as medicine, a long-run endeavor has been made to build knowledge and semantic databases: numerous nomenclatures and thesauri are being used in medical information processing, from patient data classification for statistical purposes to decision-support (Musen and van bemmel, 1997).
There is nevertheless a constant need to extend and tune existing technical vocabularies to account for new words and new word usages. The constant changes in techniques and approaches as well as variations in terminology which can be observed for the same specialty in different places (Hersh et al., 1997) prevent these fundamental resources from being complete. To address this problem, many works try to combine semantic resources and corpora. One can either exploit a large-coverage but general semantic lexicon and tune it to specialized domains (Basili et al., 1997) or use a specific but small lexicon and try to extend it. For many languages, such as French, large semantic databases are not yet available, and the second strategy is the only possible one.
In previous work (Nazarenko et al., 1997), we explained how we use the syntactic distribution of words in a corpus, as obtained by a natural language parser, in order to get similarities between words. We ran several experiments with ZELLIG (Habert et al., 1996), a suite of NLP tools, on the corpus gathered for the European project MENELAS (Zweigenbaum and Consortium MENELAS, 1994) in the field of coronary diseases, using high-level SNOMED axes (Cot et al., 1993) as comparison categories. In this paper, we attempt to quantify the extent to which, given a core lexicon and a specialized corpus, it is possible to assign a correct semantic category to unknown words by relying on their syntactic distribution and comparing it to that of already categorized words.
We first shortly recall the ZELLIG method which relies on syntactic distribution to build a similarity map of the corpus words. Then we present our proposal for categorizing unknown words according to the "votes" of their immediate neighbors. We describe the parameters which drive the categorization process, and study the actual influence of these parameters. Finally, we discuss the biases of our method and the experiments we plan.

2 Computing Syntax-Based Similarities

ZELLIG (Habert et al., 1996) relies on normalized syntactic nouns phrases (NPs) as local contexts for words. It uses parse trees retrieved by NP extractors: in the present experiment, LEXTER (Bourigault, 1993). ZELLIG automatically reduces the numerous and complex noun phrases provided by LEXTER to *elementary dependency trees*, as these normalized NPs more readily exhibit the fundamental binary relations between content words. For instance, from the parse tree α in figure 1 for sténose serrée du tronc commun gauche (*tight stenosis of left common mainstem*), ZELLIG yields the set of elementary trees *a*, *b*, *c*, and *d*. Note that trees *a* and *c* correspond to contiguous words in the original sequence, whereas *b* and *d* do not appear as such but exhibit nevertheless the dependency relationships observed in the source parse tree α.
The similarity between words depends then on the amount of normalized contexts they share. For instance, the following words can replace tronc in tree *b*: allure, artère (10 occurrences), branche (3 occurrences), etc. All these words can appear in the same context: a N P N tree, whose first noun is sténose.
In order to exhibit salient similarities, a graph is computed by ZELLIG, like the one in figure 2. 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. These shared contexts are given, for some edges only, in figure 2

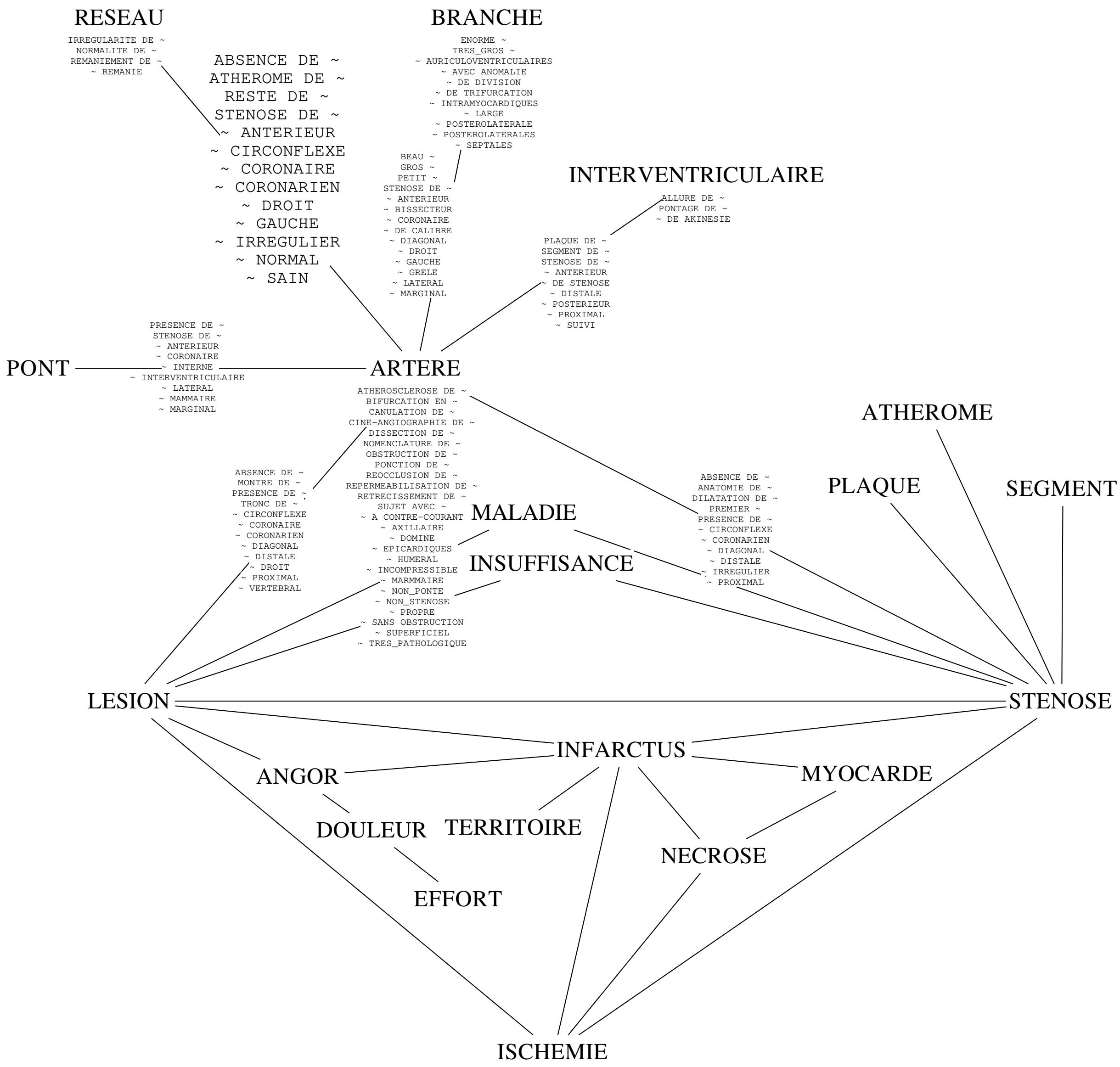

Figure 2: Graph computed by ZELLIG.

(the tilde stands for the nodes). For some nodes (réseau, branche, artère), the contexts which are not shared at all are given under the node.

Figure 3 shows the immediate edges around the word artère at threshold 9 (CTXT, weighted; see below). Each edge is labelled with the number of context types shared by the two words it links. The nodes sténose, réseau, pont, lésion, branche, and interventriculaire are the *neighbors* of artère with respect to this threshold and to this measure. If we keep this measure, but use a threshold of 10 shared contexts, interventriculaire and pont do not belong any longer to the neighbors of artère. Categories from SNOMED (/T, /M) have been added (projected) to nodes, as we explain now.

3 Using the Graph of Syntactic Similarities to Categorize Unkown Words

Assuming that graph edges represent similarities between words, our hypothesis is that given a (supposedly) unknown word, its semantic category can be determined as the most salient among that of its neighbors. The present experiment tries several formulas to compute this salience. To assess the performance of the method, we prepared a test set in which the words of the MENELAS corpus were categorized according to the high-level axes of the SNOMED International nomenclature (Cot et al., 1993): 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**). We chose this vocabulary as a compromise between clinical coverage (Chute et al., 1996) and availability in French: the SNOMED Microglossary for Pathology (12,500 terms) has a French version (Ct,

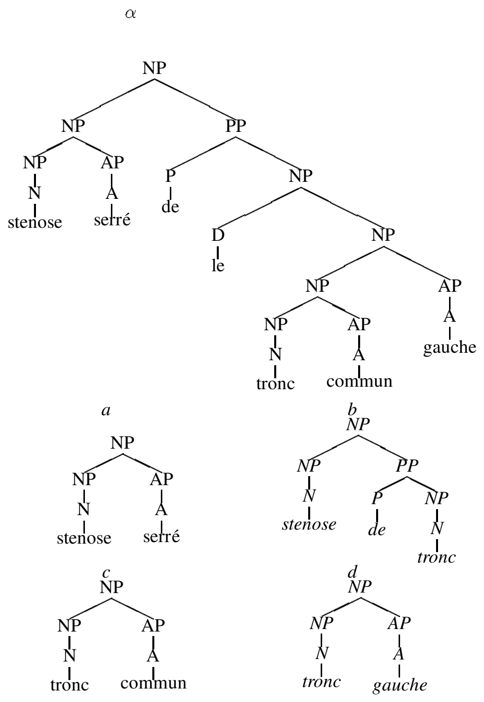

Figure 1: Complete parse tree (α) and corresponding elementary dependencies (*a*, *b*, *c*, *d*).

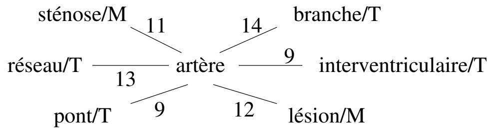

Figure 3: Edges around the word artère.

1996). The NPs extracted by LEXTER include 2,426 different lemmas. We categorized, mainly manually, the 2,073 lemmas occurring in the elementary trees of the corpus, to obtain a base semantic lexicon. We projected these categories onto the words on the nodes of the graph. We then had a procedure examine each word in turn, assuming its category were unknown, guessing it from its neighbors, and comparing the guess with the reference category found in the lexicon. For instance, in figure 3, we knew beforehand that artère belongs to category **T**. However, we erased that category, thus creating artificially an unknown word.
To tune our method, we varied several parameters: the similarity measure in the graph construction, the similarity threshold that prunes the graph and decides whether two words are neighbors, and the formula for choosing the category of a word among the categories of its neighbors. The remainder of this section details similarity computations.

3.1 Computing a similarity value

Each word W_i occurs in a set of different context types $\{ctx_i^k\}$, each with a number of occurrences $\left|ctx_i^k\right|$. 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 use in the present experiment several similarity measures:

1. their number of shared context types: $sim\text{-CTXT}_{ij} = \left|\{ctx_i^k\} \cap \{ctx_j^k\}\right|$
2. the frequency of their shared context types: $sim\text{-CTXO}_{ij} =$ $\sum_{k \in \{ctx_i^k\} \cap \{ctx_j^k\}} \min(\left|ctx_i^k\right|, \left|ctx_j^k\right|)$
3. a Jaccard measure (Saporta, 1990, p. 242) over context types:

   $sim\text{-JACCARD}_{ij} = \frac{sim\text{-CTXT}_{ij}}{\left|\{ctx_i^k\} \cup \{ctx_j^k\}\right|}$.

   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*-CTXT$_{ij}$), but also in many distinct contexts.

In the original method (Habert et al., 1996), two words are considered similar if they share enough (according to the choosen threshold) *types* of contexts. This is the rationale behind the first similarity measure above (CTXT). The frequency of each context as well as its number of occurrences are not taken into account. The second measure (CTXO) uses the number of shared occurrences as well as the number of shared types of contexts. According to the third measure (JACCARD), two words are very similar if they share most of their contexts.

3.2 Pruning the graph

Given a measure, we first prune the edge for each couple of words whose similarity falls below a given threshold. For instance, for CTXT, sténose has 317 neighbors at threshold 2, 23 at threshold 5, and only 2 at threshold 15 (infarctus and lésion). We examined a relevant series of thresholds for each kind of measure: $\{1\ 2 \ldots 19\}$ 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.

3.3 Ranking the categories of the immediate neighbors

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

1. each vote is counted as one (unweighted);
2. the vote of each neighbor W_j is weighted by its similarity *sim*-X_{ij} with the word W_i, the more similar words having higher weights.

Votes for the same category are summed, and categories are ranked according to their cumulated votes. The top-ranked category is hypothesized to be the category of the word W_i. For instance, on figure 3, with W_i = artère, in the weighted agregation scheme, $\sum_{\text{cat}(j)=T} sim\text{-CTXT}_{ij} = 45$ and $\sum_{\text{cat}(j)=M} sim\text{-CTXT}_{ij} = 23$ (in the unweighted agregation scheme, $\sum_{\text{cat}(j)=T} sim\text{-CTXT}_{ij} = 4$ and $\sum_{\text{cat}(j)=M} sim\text{-CTXT}_{ij} = 2$), so that category T is (correctly) ranked first.

4 Results

We examine results as follows. Given a word, our procedure ranks the categories of the neighbors, from most to

CTXT					CTXO					JACCARD				
	unweighted		weighted			unweighted		weighted			unweighted		weighted	
thr	pre	rec	pre	rec	thr	pre	rec	pre	rec	thr	pre	rec	pre	rec
1	45.8	32.3	51.3	36.2	1	45.8	32.3	53.0	37.3	0.00	45.8	32.3	51.7	36.4
2	49.5	13.7	56.0	15.5	2	48.0	16.6	57.1	19.8	0.02	46.5	32.6	51.7	36.4
3	55.7	8.5	65.1	9.9	3	51.2	12.4	59.6	14.4	0.04	47.4	32.8	52.6	36.4
4	56.7	5.1	66.8	6.0	4	50.8	9.1	63.4	11.3	0.06	47.7	32.6	52.3	35.7
5	60.4	3.8	69.5	4.4	5	48.9	7.3	61.6	9.2	0.10	46.6	30.3	51.2	33.3
6	60.5	2.7	67.0	3.0	7	47.3	5.1	62.2	6.7	0.15	45.7	25.6	50.2	28.2
7	57.5	1.9	68.8	2.3	10	51.4	3.7	58.9	4.2	0.20	44.4	22.6	49.1	25.0
8	66.1	1.5	75.0	1.7	15	44.7	1.9	55.3	2.3	0.30	43.4	16.0	47.1	17.3
9	70.5	1.3	75.0	1.4	20	47.8	1.4	55.1	1.6	0.40	44.6	12.7	45.9	13.0
10	64.1	1.0	69.2	1.1	...					0.50	43.6	11.9	45.0	12.3
11	55.2	0.7	65.5	0.8						0.60	43.4	6.8	43.4	6.8
12	47.8	0.5	56.5	0.5						...				
13	33.3	0.2	44.4	0.3										
...														

Table 1: Precision (pre) and recall (rec) figures (in %) for the different parameters (thr = threshold).

least salient. The correct category for the word may get ranked first (1), which is the desirable situation. It may also be ranked second (2), third (3), etc., or may not be present at all in the neighbors (0). There may also be a tie between the first and following categories (1-). We examine the distribution of ranks among the words of the graph obtained at the various thresholds for each measure (CTXT, CTXO, JACCARD) and 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 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.*, 2426 words). Table 1 provides precision and recall figures for the six combinations of similarity and weighting and for the main thresholds.

5 Discussion

5.1 Variation of precision and recall with parameter settings

Weighting significantly (by 5 to 10 %) increases both precision and recall for all methods and at most thresholds. **Precision** varies with the **threshold** for CTXT: it raises from 46/51 % (depending on whether weighting is off/on) at threshold 1 to an optimum of 70/75 % at thresholds 8–9, and then decreases. CTXO displays slightly less variation, while JACCARD is remarkably stable around 44–47/49–52 % at thresholds 0.00–0.20 and keeps a 43 % precision at the maximum threshold. **Recall** quickly decreases with the **threshold** for CTXT and slightly less for CTXO, while the decrease for JACCARD is much slower.

In summary, weighting seems to be beneficial in all cases. Maximum recall requires to use a low threshold for all methods (except for JACCARD, but using a higher threshold does not increase precision in that case). At low thresholds, the three methods do not display significantly different precisions.

With CTXT, which displays significantly 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 tresholds 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.

On the whole, among the 2426 words in the NP corpus, 1711 (70.5 %) appear in the threshold-1 CTXT graph, the missing words being hapaxes (words with a single occurrence in the corpus) 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 %.

5.2 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 to 65.4 % for G (the category S for which no correct category was found does not have a significant size; the 12 J words in our corpus happened to all be hapaxes with specific contexts). Table 2 shows that the categorization precision is roughly correlated with the number of words in each category, the largest categories G and F obtaining the highest precision.

Assigning a probability to each category according to its frequency gives a better baseline for comparison. Table 2 shows that, for all categories, the actual categorization precision is significantly higher than the probability score.

On the contrary, the categorization results are not correlated with the graph density. 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.

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. It can also bring out a minor category, the second most salient one. For instance, among the G-

Category	Nb of words	Correctly categorized Nb	%	Baseline
A	12	2	16.7 %	1.8 %
C	8	3	37.5 %	1.2 %
D	24	4	16.7 %	3.6 %
F	110	61	55.4 %	16.4 %
G	301	197	65.4 %	44.8 %
L	7	1	14.3 %	1 %
M	58	10	17.2 %	8.6 %
P	63	22	34.9 %	9.4 %
S	2	0	0 %	0.3 %
T	87	33	37.9 %	12.9 %
J	0	0	0 %	0 %
Total	672	333	49.6 %	26.3 %

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

categorized words, one can contrast the G_T (G + topography) and G_P (G + procedure) subgroups:

G_T antérieur antéro-apical apical collateral minime significatif sévère ...

G_P actuel année immediat jour mois possibilité réalisation réapparition ...

which may prove relevant for further subcategorization.

5.3 Limitations

The analysis of erroneous categories brings out the major cases of categorization ambiguity. 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. Note that Hirschman et al. (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, there is a discrepancy concerning the granularity of the vocabulary. SNOMED entries are mainly multi-word terms. However, in order to project this nomenclature onto the words of the graph, we had to work on the components (single words) of these terms. As a result, the head word of each term, which often represents its hypernym, is 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 primarily affects recall.

Last, the morpho-syntactic tagger used by LEXTER is not error-free. Tagging errors imply erroneous attachments and phrases, which affects both precision and recall.

6 Perspectives

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.

This categorization method, originally developed for tuning an incomplete nomenclature for a given technical corpus, can have various other applications.

It could be used for progressively enriching 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 as 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 from place to place and time to time, it may be difficult to manually identify the category of an unknown word which could be a "faux ami" 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 saw how the voting results can help to subcategorize a general category such as the SNOMED G axis. However, such an application requires to test our method on larger corpora. The fact that a method is well suited for moderate-size corpora and even for additional texts does not imply that it is equally suited for larger ones.

Acknowledgements

We thank Dr. RA Côté for graciously providing us an early copy of the French version of the SNOMED Microglossary for Pathology.

References

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

Bourigault, Didier (1993). An Endogeneous Corpus-Based Method for Structural Noun Phrase Disambiguation. In *Proceeding 6th EACL* (pp. 81–86). Utrecht.

Chute, Christopher G., Cohn, Simon P., Campbell, Keith E., Oliver, Diane E. & Campbell, James R. (1996). The Content Coverage of Clinical Classifications. *J Am Med Informatics Assoc*, vol. 3 (3), 224–233.

R. A. Côté, D. J. Rothwell, J. L. Palotay, R. S. Beckett & L. Brochu, editors (1993). *The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International*. Northfield:College of American Pathologists.

Côté, Roger A (1996. *Répertoire d'anatomopathologie de la SNOMED internationale, v3.4*. Université de Sherbrooke (Sherbrooke, Québec.

Habert, Benoît, Naulleau, Elie & Nazarenko, Adeline (1996). Symbolic word clustering for medium-size corpora. In *Proc. 16th COLING* (pp. 490–495). Copenhagen.

Hersh, William R., Campbell, Emily H. & Malveau, Susan E. (1997). Assessing the Feasibility of Large-Scale Natural Language Processing in a Corpus of Ordinary Medical Records: A Lexical Analysis. *J Am Med Informatics Assoc*, vol. 4 (suppl), 580–584.

Hirschman, Lynette, Grishman, Ralph & Sager, Naomi (1975). Grammatically-Based Automatic Word Class Formation. *Inform Proc Management*, vol. 11, 39–57.

Musen, Mark A. & van Bemmel, Jan H. (1997). *Handbook of Medical Informatics*:Springer-Verlag.

Nazarenko, Adeline, Zweigenbaum, Pierre, Bouaud, Jacques & Habert, Benoît (1997). Corpus-Based Identification and Refinement of Semantic Classes. *J Am Med Informatics Assoc*, vol. 4 (suppl), 585–589.

Saporta, Gilbert (1990). *Probabilités, analyse des données et statistique*. Paris:Technip.

Zweigenbaum, P. & Consortium MENELAS (1994). MENELAS: an Access System for Medical Records using Natural Language. *Computer Methods and Programs in Biomedicine*, vol. 45, 117–120.