Tuning an Existing Nomenclature for Specific Domain Corpora: A Syntax-Based Similarity Method

P. Zweigenbaum, Ph.D.[1] B. Habert, Ph.D.[2] A. Nazarenko, Ph.D.[3] J. Bouaud, Ph.D.[1]

[1] DIAM — Service d'Informatique Médicale/AP-HP & Dépt. Biomathématiques U. Paris 6

[2] UMR 9952 — École Normale Supérieure de Fontenay St Cloud

[3] Laboratoire d'Informatique de Paris-Nord — Université Paris 13

BACKGROUND

There is a constant need to extend and tune medical vocabularies to account for new words and new word usages. Robust natural language processing (NLP) tools can be applied to medical texts corpora such as patient narratives and help collect and analyze unknown words[1,2]. The aim of the present work is to assess the potential for classifying unknown words based on the semantic categories of "neighbors" identified through syntactic distributional properties[3].

METHODS

We worked with ZELLIG, a suite of NLP tools, on the (84 Kword) corpus gathered for the European project MENELAS in the domain of coronary diseases, using the high-level SNOMED axes as categories. ZELLIG uses parse trees retrieved by noun phrases extractors, and reduces them to elementary dependency trees. Second-order affinities show which words share the same contexts. As a third-order technique to exhibit salient similarities, a graph is computed by ZELLIG. The words constitute the nodes. An edge corresponds to a certain amount of shared contexts, according to a given measure and a chosen threshold. 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.

RESULTS

We attempted 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 similarity measure, thresholds used to prune the graph, and the vote aggregation methods for ranking the categories of the immediate neighbors. With the currently examined parameters, 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. More information on the precise experiments and their results is provided in Habert et al.[4].

CONCLUSIONS

Whereas weighting was useful, using a threshold was not really desirable, and the different similarity measures tested did not bring drastic changes at low thresholds. There is still room however for other experimentations[4]. Categorization results are significantly above chance, but not sufficient for a fully-automated process. We argue nevertheless that an automatic categorization process is a necessary tool to help a human expert. 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. 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.

References

1. Hersh WR, Campbell EH, Evans DA, and Brownlow ND. Empirical, automated vocabulary discovery using large text corpora and advanced natural language processing tools. *J Am Med Informatics Assoc* 1996;3(suppl):159–63.
2. Nazarenko A, Zweigenbaum P, Bouaud J, and Habert B. Corpus-based identification and refinement of semantic classes. *J Am Med Informatics Assoc* 1997;4(suppl):585–9.
3. Hirschman L, Grishman R, and Sager N. Grammatically-based automatic word class formation. *Inform Proc Management* 1975;11:39–57.
4. Habert B, Nazarenko A, Zweigenbaum P, and Bouaud J. Extending an existing specialized semantic lexicon. In: Rubio A, Gallardo N, Castro R, and Tejada A, eds, First International Conference on Language Resources and Evaluation, Granada. 1998:663–8. also available at url `http://www.biomath.jussieu.fr/~pz/biblio-pierre/#Habert:LREC98`.