Automatic indexing of health resources in French with a controlled vocabulary for the CISMeF catalogue: a preliminary study

Aurélie NEVEOL [a, b], Alexandrina ROGOZAN [b], Stéfan DARMONI [a, b]

[a] *CISMeF, Rouen University Hospital, France & L@STICS, Rouen Medical School, France*
[b] *Perception System Information Lab, FRE 2645, CNRS, INSA Rouen & Rouen University, France*

Abstract

The profusion of online resources calls for tools and methods to help Internet users find precisely what they are looking for. Quality controlled gateway CISMeF provides such services for health resources. However, the human cost of the documentary tasks involved in maintaining and updating the catalogue are increasingly high. This paper presents the automatic indexing system currently developed in the CISMeF team to assist human indexers. The results of a preliminary evaluation indicate that the automatic indexing strategy is relevant. Moreover, the system presented in this paper retrieves keyword/qualifier pairs as opposed to single terms. Further development and tests will be carried out before the efficiency of the system can be assessed.

Keywords:

Controlled Vocabulary, Automatic Indexing, MeSH, Natural Language Processing.

Objective

Our research work aims to develop an automatic indexing system that would broaden the CISMeF catalogue coverage (13,000 resources on 01/01/04) while ensuring good indexing quality according to the current manual criteria. At first, this system will be used as an indexing help tool for the librarians, so as to reduce the manual indexing delays while assessing the quality of the automatic indexing produced. Whereas existing MeSH extractors are able to retrieve keywords and qualifiers separately, we shall focus our efforts on the extraction of keyword/qualifier pairs from the resources. The use of this novel feature is an important step towards accurate indexing.

Automatic Indexing Strategy

Each resource is indexed with a list of terms (keywords or keyword/qualifier pairs, which is more precise) taken from the MeSH (Medical Subject Headings). Based on the manual procedure, our automatic indexing strategy consists in:

1. Locating textual elements in the resource
2. Mapping these elements to MeSH keywords (or pairs)
3. Using terminology properties regarding keywords hierarchy, keyword/qualifier associations and *check tags*.
4. Computing a score (tf*idf , text length)
5. Selecting the final index with a breakage function
6. Major / Minor weighting for selected keywords (pairs)

French being a morphologically rich language, we have chosen to abide the consensus in the indexing community and to locate textual elements in the form of MeSH terms, inflected or derived forms of MeSH terms, MeSH synonyms and inflected or derived forms of MeSH synonyms. The very large size of the dictionaries thus involved lead us to use automata for the detection of textual elements. This kind of technique is implemented in the linguistic platform INTEX, which we decided to integrate in our system. Several indexing rules, consisting in the mapping of recurring expressions to keyword/qualifier pairs have been provided by the chief librarian in charge of superindexing in the CISMeF catalogue and are implemented with transducers. In addition to being an effective method complexity-wise, the use of automata and transducers for term extraction ensures that noise will be very limited, which is our chief concern. On the other hand, since it is impossible to provide an exhaustive list of all MeSH terms variations, some silence is to be expected.

A preliminary experiment was conducted on a sample of 10 resources extracted from the CISMeF catalogue as addressing diabetes related topics. The dictionaries used covered check tags, qualifiers, and diabetes related keywords. The automatic indexing produced by the system for each resource was compared to the manual indexing available in the catalogue. The preliminary results show that the general indexing strategy is relevant, and more specifically that the coverage of the dictionary on terms related to diabetes is good. The system achieves a precision that is comparable to that of other existing operational systems. Several improvements, including significant MeSH dictionary coverage increase and score computation adjustments have to be carried out before this tool can be pronounced practically effective.