Available online at www.sciencedirect.com

SCIENCE DIRECT®

Information Processing and Management 42 (2006) 695–709

INFORMATION PROCESSING & MANAGEMENT

www.elsevier.com/locate/infoproman

Automatic indexing of online health resources for a French quality controlled gateway

Aurélie Névéol [a,b,*], Alexandrina Rogozan [a], Stéfan Darmoni [a,b]

[a] *PSI Laboratory, CNRS FRE 2645, INSA de Rouen, Place Emile Blondel, BP 08, 76131 Mont-Saint-Aignan Cedex, France*
[b] *CISMeF, Rouen University Hospital, and L@STICS, Rouen Medical School, 1, rue de Germont 76031, Rouen, France*

Received 19 October 2004; accepted 5 January 2005
Available online 2 March 2005

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 maintaining and updating the catalogue are increasingly high. This paper presents the automatic indexing system currently developed in the CISMeF team to be used as such for preliminary indexing, or after human reviewing for the final indexing. The system architecture, using the INTEX platform for MeSH term extraction is detailed. The results of a first evaluation tend to indicate that the automatic indexing strategy is relevant, as it achieves a precision comparable to that of other existing operational systems. Moreover, the system presented in this paper retrieves keyword/qualifier pairs as opposed to single terms, therefore providing a significantly more precise indexing. Further development and tests will be carried out in order to improve the coverage of the dictionaries, and validate the efficiency of the system in the indexers' everyday work.

Keywords: Automatic indexing; Controlled vocabulary; Corpus analysis

1. Introduction

Internet has become a very prosperous source of information in numerous fields, including health. The CISMeF project (French acronym of Catalogue and Index of Medical On-Line Resources) described in

[*] Corresponding author. Address: PSI Laboratory, CNRS FRE 2645, INSA de Rouen, Place Emile Blondel, BP 08, 76131 Mont-Saint-Aignan Cedex, France. Tel.: +33 2 3288 8800x61119; fax: +33 2 3288 8832.

E-mail addresses: aneveol@insa-rouen.fr (A. Névéol), arogozan@insa-rouen.fr (A. Rogozan), stefan.darmoni@chu-rouen.fr (S. Darmoni).

0306-4573/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2005.01.003

Darmoni et al. (2000) was initiated in 1995 in order to meet the users' need to find precisely what they are looking for among the numerous health documents available online. As a Quality Controlled Health Gateway (Koch, 2000), CISMeF describes and indexes the most important resources of institutional health information in French. It currently contains more than 14,000 resources, and it is updated manually with 50 new resources each week. Indexing is a decisive step for the efficiency of information retrieval within the CISMeF catalogue, and it is also one of the most time consuming tasks for the librarians, demanding high-level skills. However, an ever increasing number of documents, be it articles, clinical guidelines or teaching material become available on electronic form, and there is not enough time to index them as they are published. For teaching resources only, although 3000 have already been indexed by the CISMeF indexers, 1500 additional resources are waiting to be indexed, and included in the catalogue. If we take these figures as representative of the situation for all resource types, we can assume that a total of 6500 resources can not be included in the CISMeF catalogue yet because of indexing delays.

Therefore, it is necessary to create automatic systems for indexing purposes. Several researchers have addressed this issue for the general domain with FASTR (Jacquemin & Royauté, 1994) for instance. Other projects were more specifically dedicated to the medical field, such as the Indexing Aid Project (Humphrey & Miller, 1987), the NLM (National Library of Medicine) Indexing Initiative (Aronson, Mork, Gay, Humphrey, & Rogers, 2004), the HONselect (Gaudinat & Boyer, 2002) and MeSHMap (Ruch, Baud, & Geissbühler, 2003) systems, and NOMINDEX (Pouliquen, 2002) to name but a few.

2. Objective

This work aims to develop an automatic indexing system that would help broaden the CISMeF catalogue coverage while ensuring good indexing quality. The same system will be used as a fully automatic process (without any human revision) to produce a temporary indexing for pending resources and as a semi-automatic tool to produce the final indexing. In this case, the indexing produced by the automatic system will be revised by a human indexer. The ultimate goal is to reduce the indexing delays by replacing the current fully manual indexing by semi-automatic indexing.

In fully automatic mode, the system performance is evaluated by comparison to manual indexing which stands as the reference indexing. In the evaluation of the NOMINDEX system conducted in 2001 on a representative sample of CISMeF resources (Pouliquen, 2002), the noise rate was found much too high by the indexers who reckoned revising the automatic indexing of the system took much longer than producing a fully manual indexing from scratch. This was partly due to the fact that NOMINDEX was not fully adapted to the CISMeF indexing standards. Therefore, the new system should be compliant with the current CISMeF manual indexing criteria, and keeping the noise rate as low as possible will be a chief concern, as long as the most relevant keywords are retrieved from a resource.

Our goal is to succeed in producing as accurate and reliable an indexing as possible with the fully automatic system, so that little time will be required for the human revision. Hence, whereas existing automatic indexing systems in the medical field are able to retrieve keywords and qualifiers separately, we shall focus our efforts on the extraction of keyword/qualifier pairs from the resources, as this novel feature is an important step towards the most accurate indexing.

The resources to be indexed in CISMeF are very heterogeneous in terms of content type (guidelines, patient information,...) and size. In the experiment presented in this paper, we have selected a collection of representative diabetes related resources which size ranged from a couple of pages for patient information to 100 pages for guidelines.

3. Indexing health resources

3.1. CISMeF manual indexing policy

The 14,000 resources currently referenced in CISMeF have been indexed manually by five medical indexers according to very specific criteria, based on those used in the Medline database, and described in (Dailland, Leuthereau, & Vallée, 2003).

In order to build a system compliant with all CISMeF manual criteria, we first focused on producing a thorough description of the indexing policies and practices as recommended by Milstead (1992).

Each resource is indexed with a list of terms (keywords or keyword/qualifier pairs) taken from the MeSH (Medical Subject Headings), which is the reference thesaurus developed by the US NLM for the bio-medical domain. As CISMeF deals with resources in French, the French translation of the MeSH produced and maintained by the French Medlars Center, the National Institute for Health and Medical Research (INSERM) is used.[1]

The MeSH 2004 contains approximately 22,600 hierarchically arranged keywords and 84 qualifiers that can be coordinated to the keywords, in order to refer to particular aspects of a subject. Hence, it is more accurate to produce an indexing based on keyword/qualifier associations, rather than single terms. For instance, if a resource details the various methods used for the diagnosis of a given disease, it is more relevant to index this resource with the pair <*disease D*/*diagnosis*> than with the single keyword <*disease D*>. This way, users who need other information related to disease D will know that they have to look in another resource.

In order to be more accurate in describing the resource that is being indexed, a major weight is allocated by the CISMeF indexers to keywords (or keyword/qualifier pairs) representing a concept that is detailed or dealt with in the whole article. A minor weight will be allocated to the keywords (or pairs) representing a concept that is only mentioned or treated in a small section. For instance, if a resource discusses nutrition for patients with type 1 diabetes, <*diabetes mellitus*, *type I*> and <*diabetic diet*> would be major keywords as <*diabetic diet*> is the main, central topic of the resource, and <*diabetes mellitus*, *type I*> is the specific context of the statement. On the other hand, <*energy intake*> or <*diet*, *fat-restricted*> would be minor keywords if the concept they represent are discussed at some point, but they are not the most detailed topics of the resource. In summary, major keywords should be the ones that spring to mind as an answer to the question: in a nutshell, what is the resource about? Minor keywords complement this description of resource content.

Indexing also takes check tags into account. Check tags are MeSH keywords singled out as priority indexing terms, to remind indexers to include information about patients or subjects of an experiment when mentioned in the resource. CISMeF uses all the human medicine related check tags.

Furthermore, the size of the index also referred to as the exhaustivity variable by Anderson and Pérez-Carballo (2001) is different for every resource. Depending on how dense a resource content is, or how detailed a description the user would expect for a given resource, the indexer chooses to use few or many keywords (or pairs) to describe it. This practice reflects the findings of Soergel (1985) and Lancaster (1998) who insist on the user-oriented aspect of the indexing task. For instance, few keywords (or pairs) are really necessary to describe a hospital website, as one already knows what to expect from this type of resource: a description of the facilities, location, contact details for the various health-care departments, and so on. If special features such as medical articles are also offered, they may be indexed independently. On the other hand, up to several dozens keywords (or pairs) may be required to index a clinical guideline. In fact, in addition to being usually quite lengthy (up to a few hundred pages), these resources are systematically developed

[1] cf. http://disc.vjf.inserm.fr:2010/basimesh/mesh.html.

statements designed to help practitioners and patients decide on appropriate healthcare for specific clinical conditions and/or circumstances. Each particular condition or circumstance and the corresponding recommended steps for diagnosis and adapted treatment are detailed. Hence, the indexing must reflect it by using the keywords (or pairs) relevant to every case, so that users looking for information on one of these conditions or circumstance shall be able to retrieve this specific resource. Although long resources are likely to be denser in content than shorter ones, it must be noted that resource length is not the only variable influencing the size of the index. It is not possible to set a fixed exhaustivity rate such as "1 keyword (or pair) per 300 words" since the number of topics discussed in any three-hundred-word resource is different from one resource to another. In summary, the size of the index depends mainly on the resource type, the length of the resource, and the number of topics covered in the resource.

All the features described above have to be modelled into the indexing system. We are now going to detail how we plan to achieve this in our experimental system.

3.2. CISMeF automatic indexing system

"[T]here is a two step process in human indexing: (1) the analysis of a text, (...) and (2) the translation of this notion (...) into the indexing language (...). Most of the rules regarding indexing, cataloguing, and classification relate to the second step".

Anderson and Pérez-Carbollo's (2001, p. 247) description of human indexing sums up the steps that were modelled in the CISMeF Indexing System. After an overview of the system architecture (Section 3.2.1), we give a detailed presentation of the indexing steps:

Resource analysis consists in locating textual elements related to the MeSH keywords and qualifiers (Section 3.2.2).

Translation into MeSH headings and subheadings is performed by exploiting dictionary information (such as lemmatisation, synonym relations, etc.) and indexing rules (Section 3.2.3).

Additional rules enable further refinement of the resulting indexing (Sections 3.2.4 and 3.2.5).

3.2.1. Overview of the system

As shown in Fig. 1, the CISMeF Indexing System interacts with the linguistic platform INTEX developed by Silberztein (1993). INTEX is a powerful corpus analysis tool, which may also be used as a linguistic toolbox, by integrating calls to INTEX functions in an outside application. Two types of databases are used for the automatic indexing: a MeSH dictionary, and three knowledge bases containing information on the

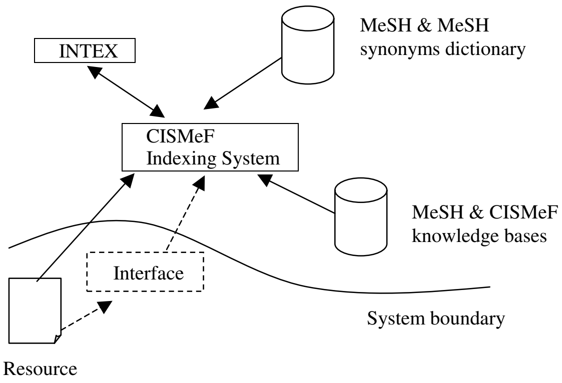
INTEX
MeSH & MeSH synonyms dictionary
CISMeF Indexing System
MeSH & CISMeF knowledge bases
Interface
System boundary
Resource

Fig. 1. System architecture.

MeSH hierarchy, the MeSH check tags, and the keyword/qualifier association history drawn from previous CISMeF notices. The development of the indexing system contributed to enrich these databases (Névéol, Rogozan, Douyère, & Darmoni, in press), and to formalize some indexing practices as recommended by Milstead (1992).

3.2.2. Identification of textual elements

It must be noted that some "global" keywords cannot be retrieved thanks to cues from the resource, or else at great cost. For example, if a resource presents the results from a study conducted in France, and from similar studies conducted in Germany and Sweden, the appropriate MeSH keyword would be the "global" keyword <*comparative study*>. Although this reasoning is very basic for a human indexer, it proves more difficult to have it performed automatically. However, most keywords can still be inferred from textual elements within the resource. Accordingly, the first step of the indexing algorithm consists in the identification of textual elements considered as useful for the indexing, which are listed in a comprehensive MeSH dictionary. These textual elements may be MeSH terms (i.e. MeSH keywords or qualifiers) such as <*sujet âgé*> (<*aged*>), inflected MeSH terms such as "sujets âgés" (plural form of <*sujet âgé*>), synonyms of MeSH terms such as "personne agée"("elderly") or inflected synonyms of MeSH terms such as "personnes agées"(plural form of "personne agée").

The very large size of the dictionaries involved lead us to use automata for the detection of textual elements. In fact, automata may be used for the detection of dictionary entries in a text by going through the text only once. Moreover, Crochemore and Rytter (1994) show that processing time does not depend on the size of the dictionaries. This kind of technique is widely used in computational linguistics and it is implemented in INTEX (Silberztein, 1993), which we decided to integrate in our system. INTEX deals with the DELA-type dictionaries introduced by Courtois and Silberztein (1990).

Several indexing rules have been provided by the indexing expert in charge of supervising the indexing in the CISMeF catalogue. These rules consist in extracting keyword/qualifier pairs from recurring expressions. For instance, the pair <DISEASE D/*Prevention* & *Control*> should be deduced from the French expression "vaccin contre la MALADIE M" ("vaccination against DISEASE D"), where we define DISEASE D as a MeSH keyword belonging to the C or F03 MeSH trees. They are implemented in the system as local grammars, in the form of INTEX graphs. Fig. 2 shows the complete graph dedicated to the qualifier <*prevention* & *control*>, which includes the sample rule given above. Similar grammars are also used for a comprehensive extraction of age group keywords, as described by Gaudinat and Boyer (2002). At the moment, a total of ten graphs have been developed, but we are currently working with the indexing expert on the formalization of general indexing rules related to all qualifiers. To our knowledge, this type of work has not been attempted before. We are proceeding as follows:

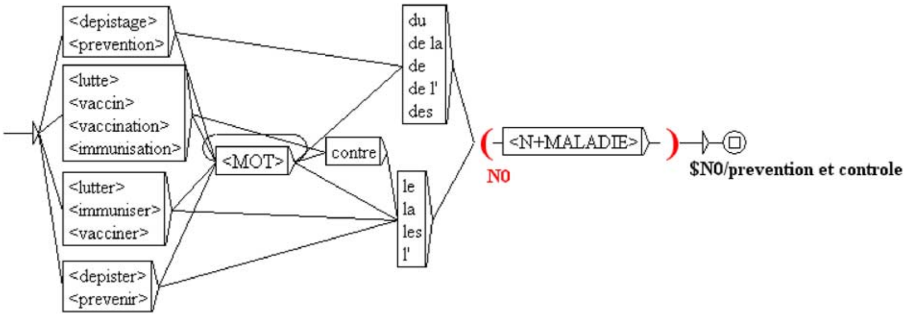

Fig. 2. Local grammar recognising the keyword/qualifier pair disease D/prevention & control in a French text.

- In an interview, the knowledge engineer (AN) helps the indexing expert to describe how the use of a particular qualifier arises. Together, they look through precise examples (drawn from the CISMeF database) where the qualifier has been used for the indexing of a resource. They try to identify a pattern in the resources, or a recurring situation leading to the indexing process.
- The knowledge engineer analyses the interview material and writes a formal summary of the indexing rules that were described, to be validated by the indexing expert.

The knowledge engineer finds a technical solution to effectively implement the rules in the system.

We estimate that approximatively one working day is necessary to carry out the description of one qualifier, and implement it in the system.

3.2.3. Mapping to MeSH terms and use of knowledge bases

The second step consists in mapping the textual elements to the corresponding standard MeSH keywords or qualifiers, in order to create a list of MeSH terms occurrences including their position in the resource.

The third step will exploit this information to produce the final indexing for the resource. Information on the positions of each term (keyword or qualifier) is used to infer keyword/qualifier pairs that could not be retrieved with the local grammars. Indeed, as mentioned for single MeSH terms at the beginning of the previous section, some keyword/qualifier pairs cannot be retrieved solely based on cues from the resource, and human indexers are able to come up with them after a global analysis of the resource or relevant passage. For example, a resource may discuss "nutrition education" for diabetic patients, and detail various aspects of the diet recommended for each particular condition without any explicit reference to a "diet therapy". Although human indexers will be able to summarize this type of content with the MeSH pair <*diabetes/diet therapy*>, inferring it automatically proves more difficult. In this preliminary study, we have decided to leave this thorny problem on the side, and to focus on cases where pairs may be inferred from textual elements within the resource.

Based on this assumption, we estimate in this preliminary study that a keyword and a qualifier appearing in the same phrase or sentence are very likely to be an instance of the corresponding keyword/qualifier pair. For instance, in the expression "complications aiguës du diabète insulinodépendant" ("serious complications of type 1 diabetes"), the elements "complications" corresponding to qualifier <*complications*> and "diabète insulinodépendant" corresponding to keyword <*diabetes mellitus*, *type I*> will be located, and associated to form the pair <*diabetes mellitus*, *type I/complications*> which is relevant in this context.

However, a keyword and a qualifier appearing in different paragraphs of the text are not likely to be related, unless we are dealing with a major keyword. For instance, in the sentence "Elle augmente la durée d'hospitalisation et le taux de mortalité dès le 30$^{\text{éme}}$ jour." ("Duration of hospitalisation and mortality rates are increased after the 30th day") the elements "hospitalisation" and "mortality" will be located. An attempt to form the pair <*hospitalisation/mortality*> will fail because this association is not valid according to the MeSH terminology. Hence, the system will attempt to associate the qualifier <*mortality*> with the most frequent keyword in the resource, namely <*diabetes mellitus*>. Since <*diabetes mellitus/mortality*> is a valid association, the pair will be considered as a relevant indexing term, which is in this precise case relevant.

Therefore, it seems relevant to try to associate keywords and qualifiers that have close positions within the text, and when this is not possible, associating single qualifiers with the resource most frequent keyword may be considered. The prospective pairs will then need to be checked from the list of MeSH permitted associations.

3.2.4. Score computation and selection of final index

The final indexing is based on a score S_i that is computed for each keyword (or pair) i.

According to the consensus among CISMeF indexers, check tags are almost systematically selected: their score S_i is set to a maximum value when they appear more than once. For other keywords (or pairs), the

- endocrine diseases
  - **diabetes mellitus**
    - **diabetes mellitus, type I**
    - diabetes mellitus, lipoatrophic
    - diabetes mellitus, nephrogenic
    - **diabetes mellitus, type II**

Fig. 3. Sample MeSH hierarchy relationships.

score S_i is based on the number of occurrences in the resource. Identification of a hierarchy relationship between keywords (as shown on Fig. 3) initiates an even score reallocation from the father keyword to the child(ren) keyword(s), so that only the more precise terms will remain.

Thus, if there are 10 occurrences for <*diabetes mellitus*>, 15 for <*diabetes mellitus*, *type I*>, and 4 for <*diabetes mellitus*, *type II*>, the hierarchical relationships between <*diabetes mellitus*> and <*diabetes mellitus*, *type I*> on the one hand and <*diabetes mellitus*> and <*diabetes mellitus*, *type II*> on the other hand will result in scores 15 + (10/2) = 20 for <*diabetes mellitus*, *type I*>, and 4 + (10/2) = 9 for <*diabetes mellitus*, *type II*> whereas <*diabetes mellitus*> will not appear on the keyword list anymore. According to Salton's blueprint (Salton & McGill, 1983) for indexing, we compute a tf*idf weight from the resulting occurrence number, so as to favour terms with high frequency in the resource, that are also sufficiently specific in the collection to be representative of the resource treated.

For a given resource r, the N_r index candidate terms are ordered by decreasing scores (for a given i, $S_i > S_{i+1}$) and only those ranked above an adaptive threshold T_r computed with a breakpoint function are retained. T_r is computed as follows:

$$T_r = \underset{i=1,\ldots,N_r-1}{\arg\max} \left\{ \frac{S_i - S_{i+1}}{S_i + S_{i+1}} \right\} \tag{1}$$

Table 1 details the calculation steps for a sample 4-candidate list. The resulting threshold is $T_r = 2$, which means that only the two first candidates of this list would be selected for the final index.

Such breakpoint functions are used to detect discontinuities in a given signal (Abdallah, 1998). In our case, a discontinuity in the scores should indicate a significant difference in relevance for the index terms. An experimental threshold shall also be set in order to differentiate minor and major keywords (or pairs).

To sum up the scoring process, we can say the score allocated to each keyword (or keyword qualifier pair) is a function of its number of occurrences in the resource, the CISMeF association history, the check tag list, and the MeSH hierarchy.

3.2.5. Enhancement of the final index

Another set of indexing rules (~100) is used to enhance the final index. There are two types of such post-treatment rules:

Table 1
Sample threshold calculation

Rank i	MeSH descriptors	Score S_i	$\frac{S_i - S_{i+1}}{S_i + S_{i+1}}$
1	Diabetes mellitus, type I	200	0.08
2	Hypoglycemic agents	170	0.84
3	Hepatitis A	15	0.58
4	Blood	4	–

- *NLM rules* indicating that single keywords are to be preferred to keyword qualifier pairs when they represent the same concept. For example, the pair <*heart/transplantation*> should be replaced by the single keyword <*heart transplantation*>.
- *CISMeF rules* indicating that some keywords or pairs should be used together for a more precise indexing. For example, the keyword <*appendectomy*> should be complemented by the pair <*appendicitis/ surgery*>.

This set of rules is currently being enriched by the indexers.

4. System evaluation

4.1. Evaluation measures

The system performances were evaluated with the standard measures used in information retrieval, namely precision and recall. For comparison with other research, the measures of noise (1-precision) and silence (1-recall) are also used. The *F*-measure was also computed, giving an equal weight to both precision and recall (Manning & Schütze, 1999). All measures were computed after the retrieval of each new keyword (or pair). At this stage, we evaluate the performance of the fully automatic system (without human review). The automatic indexing produced by the system is compared to the manual indexing produced for the same resource by a human indexer. This procedure evaluates how well the system is able to reproduce the current manual indexing. In the second set of evaluations, we also attempted to evaluate the silence of manual indexing on specific descriptors by revising the manual indexing with the enforcement of a neglected rule (quasi-systematic selection of Check Tags).

We have used both the Mc Nemar test (for precision at rank 1) and Sign test (for recall and precision at other ranks) in order to assess the significance of the comparisons between the different scoring methods (Tables 3–7).

4.2. Preliminary evaluations

A preliminary evaluation (Névéol, Rogozan, & Darmoni, 2004) of the automatic indexing system presented in this paper was conducted using a MeSH dictionary containing about 500 entries, including all check tags and qualifiers, the keywords related to diabetes and their synonyms. This dictionary covered about 1% of the MeSH, meaning that 1% of MeSH terms were covered by at least one entry in the dictionary. Then, a second evaluation was performed on the same set of 10 diabetes related resources (size ~ 45,500 words or ~300 KB) using much larger dictionaries (about 16,000 entries covering 33% of the MeSH), and little difference was observed regarding the noise. These experiments aimed to validate the automatic indexing procedure as a whole, and more specifically to test the method used to produce the dictionaries.

The results (~50% precision after the fourth term) showed that the general indexing strategy is relevant, and suggested that check tags may not be selected by the indexers as often as they should. Furthermore, it appeared that diabetes related high-frequency terms such as <*insulin*> or <*blood glucose*> were often wrongly selected by the automatic system, and that the comparatively low frequency of pairs (vs. single keywords) in the CISMeF catalogue tended to attribute them higher tf*idf scores.

4.3. Further evaluation

The new tests were conducted on a larger set of resources (57 resources totaling ~740,000 words or ~5 MB). The automatic indexing produced by the system for each resource was compared to the manual indexing available in the catalogue. For each resource, besides the "full text" indexing, we also produced a second indexing based on the table of contents or abstract of the resource ("TOC indexing"). Based on the previous experiments, the focus was on

- Evaluating the manual indexing silence on check tags
- Correcting the abusive selection of high frequency terms
- Assessing the use of tf*idf normalization

The manual indexing silence on check tags was evaluated by considering that the selection of a check tag by the automatic system was always relevant. For instance, if the manual indexing for a resource consists of 10 keywords (or pairs) and the system retrieves two check tags that are not included in the manual selection, we considered that there was 12 relevant keywords (or pairs) to be retrieved for this resource: the original 10 that were selected by the indexers, plus the 2 check tags retrieved by the system.

In order to correct the abusive selection of high frequency terms, we computed the probability $p(t)$ of a given term t to be selected as an index term as indicated below (2). R represents a given resource and I_R the resource index.

$$p(t) = \frac{Card\{(t \in R) \cap (t \in I_R)\}}{Card\{t \in R\}} \tag{2}$$

The probability $p(t)$ that a term t will be selected as an index term is in fact the number of times t is effectively selected over the number of times it could have been selected.

A set of 670 resources taken from the CISMeF catalogue were used to compute the probability of selection of 483 terms. For example, the probability of selecting <*insulin*> was 0.06 whereas the probability of selecting <*diabetes mellitus*, *type II*> was 0.41. Terms on which no information could be collected in the set were attributed the average probability of selection, which was 0.23. The original score was then multiplied by the term probability of selection. A similar technique was used successfully by Lahtinen (2000) to weight index terms according to their grammatical category in free text indexing.

The benefits of using tf*idf normalization (Salton & McGill, 1983) in our case were evaluated by comparing the scores resulting from tf*idf normalization and raw occurrences.

4.4. Results

As an illustration of the indexing results, Table 2 presents the indexing obtained for a sample resource with the automatic and manual indexing. The index shown in Column 1 results from the full text indexing (the TOC indexing resulted in a subset of these terms—they are shown in italic). Keywords (or pairs) are sorted by decreasing scores: the first keyword in the list was allotted the highest score, and so on. In Column 2, a star indicates the major terms. For this particular resource, after the fourth term is retrieved, three terms out of four were also selected by the human indexer (<*pregnancy*>, <*aged*> and <*hypoglycemic agents*/*administration and dosage*>) which corresponds to 75% precision (i.e. 25% noise). Similarly, a total of ten terms were expected based on the manual indexing, which corresponds to 30% recall (i.e. 70% silence) at this stage.

Tables 3–6 present the precision and recall obtained on the whole test set ($N = 57$) for full text indexing, using each of the features we described previously. As a summary of these results, Fig. 4 shows a selection of the corresponding precision/recall curves. Table 7 presents the indexing with single keywords, as opposed to indexing with pairs, shown in the other tables. In this case, a keyword is considered relevant if it was selected by the indexers alone *or* coordinated to a qualifier.

Table 2
Indexing for the resource retrieved December, 2, 2004 from http://agmed.sante.gouv.fr/htm/5/5106c.htm

Automatic indexing	Manual indexing
Pregnancy	*Diabetes mellitus, type II/drug therapy
Hypoglycemic agents/administration & dosage	*Diabetes mellitus, type II
Aged	Pregnancy
Diabetes mellitus, type II/therapeutics	Hypoglycemic agents/adverse effects
Diabetes mellitus, type II/complications	Hypoglycemic agents/administration & dosage
Hypoglycemic agents/adverse effects	Hypoglycemic agents/classification
Hypoglycemic agents	*Hypoglycemic agents/drug interactions
Diabetes mellitus, type II	Continuity of patient care
Insulin	Aged

Table 3
Indexing with pairs using tf∗idf vs. raw occurrences in the score computation

Rank	Tf∗idf	Raw occ.
	Precision–recall	**Precision–recall**
1	19.30–4.93	28.07–5.98
4	23.68–15.75	27.19–18.74
10	17.54–27.40	17.54–28.00
20	12.19–33.28	12.28–34.18
50	7.35–42.12	7.15–41.40
Threshold	**21.61–19.4 (*T* = 32)**	25.72–19.93 (***T* = 28**)

Table 4
Indexing with pairs using tf∗idf and probabilities of selection vs. raw occurrences and probabilities of selection in the score computation

Rank	Tf∗idf + *p*(*t*)	Raw occ. + *p*(*t*)
	Precision–recall	**Precision–recall**
1	19.30–5.21	24.56–7.98
4	24.56–15.70	27.63–20.54
10	16.32–26.58	17.02–28.02
20	11.23–34.54	11.05–33.40
50	7.15–42.67	6.96–41.21
Threshold	**27.02–19.98 (*T* = 5)**	**25.12–15.98 (*T* = 4)**

Table 5
Indexing with pairs using tf∗idf vs. raw occurrences in the score computation, while all check tags are considered as relevant

Rank	Tf∗idf + CT	Raw occ. + CT
	Precision–recall	**Precision–recall**
1	49.12–8.51	59.65–10.04
4	48.68–24.11	52.20–26.51
10	28.42–34.32	28.42–34.54
20	17.63–39.33	17.72–39.95
50	9.81–48.46	9.58–47.62
Threshold	**43.68–8.25 (*T* = 32)**	**48.47–28.49 (*T* = 28)**

Table 6
Indexing with pairs using tf∗idf and probabilities of selection vs. raw occurrences and probabilities of selection in the score computation, while all check tags are considered as relevant

Rank	Tf∗idf + CT + $p(t)$	Raw occ. + CT + $p(t)$
	Precision–recall	**Precision–recall**
1	49.12–8.95	63.16–12.63
4	49.56–24.26	53.07–28.46
10	27.19–33.12	27.89–34.51
20	16.67–40.28	16.49–39.21
50	9.69–48.94	9.35–47.17
Threshold	**48.65–28.16** (***T*** **= 5**)	**49.40–24.96** (***T*** **= 4**)

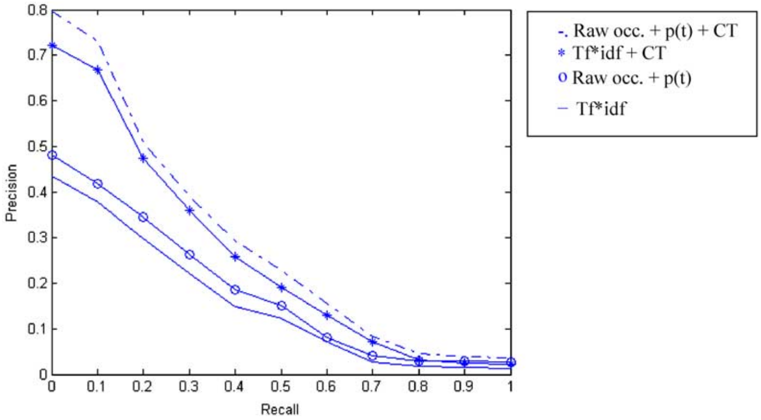

Fig. 4. Precision/recall curves.

Table 7
Indexing with single keywords using tf∗idf vs. raw occurrences in the score computation

Rank	Tf∗idf	Raw occ.
	Precision–recall	**Precision–recall**
1	42.86–11.14	28.57–7.21
4	21.43–15.96	24.11–17.87
10	15.54–25.34	15.18–23.93
20	10.18–29.71	10.36–30.89
50	6.24–35.33	6.12–34.76
Threshold	**24.16–7.43** (***T*** **= 47**)	**31.52–3.19** (***T*** **= 31**)

The last line of each table (in bold characters) shows the average rank of the adaptive threshold (between brackets, $T = *\text{average\_rank}*$) as well as the average precision and recall measures at the threshold.

For full text indexing, scores based on raw occurrences appear to give slightly better results in terms of precision and recall, especially for the first terms retrieved. The results for TOC indexing are not shown here, due to space constraints. The main difference observed was that the breakpoint function was not very efficient on TOC indexing.

According to a sign test, after rank 10, there is no significant difference between scores based on tf*idf normalization or raw occurrences. However, for highly frequent terms, it seems that raw occurrences are a better method to rank terms than tf*idf normalization based on the CISMeF collection. In fact, the difference in precision was found to be significant at rank 1 according to a Mc Nemar test ($p = 0.025$), and at rank 4 according to a sign test ($p = 0.008$). The difference in recall is not significant according to the statistical tests, but it still favors the scores based on raw occurrences.

5. Discussion

5.1. System performance

The results for full text indexing show that there is a silence of the manual indexing on check tags (Table 3 vs. Table 5). In fact, the recall rate is increased when all check tags are considered relevant. Moreover, the precision after the first terms are retrieved is also significantly higher, because our system considers it a priority to select check tags. The check tag retrieval has been assessed by the indexing expert and the system shall soon be used on all the resources in the catalogue in order to enrich the current indexing by retrieving the missing check tags.

The use of probabilities of selection induces little difference in terms of precision or recall at fixed ranks. This could mean that the set of resources used to compute the probabilities was not sufficiently large to obtain realistic figures. The most significant difference observed when using probabilities of selection concerns the adaptive threshold T, which is much lower. In fact, multiplying scores by a probability, thus a figure between 0 and 1, has the effect of reducing the scale of the scores, which leads to produce a much smaller sized index. This is actually not a bad thing, since the average number of keywords (or pairs) selected by the indexers for CISMeF resources was 4.46 in 2002, which is very similar to the threshold fixed by the breakpoint function when probabilities of selection are used (see Tables 4 and 6). Hence, the size of the automatic index seems more relevant in this case. Nonetheless, it is impossible to obtain 100% recall for all experiments when the threshold value is below the number of manually assigned keywords (which is unknown to the automatic system).

Although the system performance is better, both precision and recall are higher at the threshold than at the corresponding fixed rank, it is not always set at the point where the precision and recall rates are optimal, i.e. where the F-measure reaches a maximum.

The best results are currently obtained with full text indexing based on raw occurrences, while using probabilities of selection and assuming that the system retrieves all check tags correctly. In this case, the precision after the fourth term is retrieved is 53%, which is similar to the results obtained by HONselect (Gaudinat & Boyer, 2002).

Contrary to the other existing systems mentioned previously, our approach includes keyword/qualifier pairs in the indexing, meaning that partial retrieval is considered as non-relevant. For instance, if the system retrieves the keyword <*diabetes mellitus*/*drug therapy*> where the pair <*diabetes mellitus*/*therapeutics*> is expected, we do not consider that the system retrieved a correct indexing term. This could induce a significant precision loss as can be seen from Table 7 when tf*idf normalization is used (compare to Table 3—for example, at rank 1, the p-value was 0.005 with a Mc Nemar test). The overall precision in the retrieval of keyword/qualifier pairs in our diabetes corpus was 10.05% for all candidate pairs, and 18.31% for the final index pairs (i.e. above threshold T, as described in Section 3.2.4). Nonetheless, in the example presented in Table 2, two out of the four keywords/qualifier associations expected are effectively retrieved, and two out of the three major terms expected are also extracted by the automatic system.

It is also interesting to take into account, that, according to a study conducted by the NLM on indexing consistency (Funk, Reid, & McGoogan, 1983) human indexers usually disagree on the terms that should be selected to index a text. Consistency ranges from 76% for check tags to 33% only for keywords/qualifier pairs.

5.2. Future work

We are currently working on a benchmark evaluation of our system with the existing French MeSH indexing systems: a French version of MeSHMap (Ruch et al., 2003), HonSelect (Gaudinat & Boyer, 2002) and NOMINDEX (Pouliquen, 2002). This evaluation is performed on a corpus of randomly selected resources (all topics), and preliminary results show that 1/our system performs as well as on the diabetes related collection, and 2/matches the performance of the other systems. The final results of this comparative evaluation are to be published in 2005.

In the near future, we are planning to optimize the breakpoint function, and to enrich the lexical resources used by the system:

- *Dictionary*. We are currently working on expanding the MeSH electronic dictionary from 33% to 100% of the MeSH by March 2005. In particular, we are using NooJ (an improved version of INTEX) tools for automatic entry flexion.
- *Description of keyword/qualifier association rules*. As indicated in Section 3.2.2, we estimate that one working day is necessary to carry out the description of one qualifier. A dozen qualifiers have been covered so far. There are 84 qualifiers in total, which means about seventy days of work are needed to complete the task.
- *Further enriching of lexical resources*. We shall use the medical lexicon that is currently developed by the UMLF project (Zweigenbaum et al., 2003), and a set of association rules that were automatically extracted for information retrieval purposes (Soualmia, Barry, & Darmoni, 2003). We are also working on the automatic translation of American MeSH synonyms into French (Névéol & Ozdowska, 2005).

The automatic indexing produced by our system on a few resources to be added to the catalogue was pronounced "rather helpful" by human indexers who were asked to review it before they were added to the catalogue. Based on this positive preliminary feedback, we have decided to use the system in a production environment, and we are setting up an evaluation protocol, which will be focused on the following issues:

- *Effectiveness of the indexing system in a production environment*. Once the work on terminological resources is completed, we will launch the automatic indexing of the impending teaching resources, and collect human indexer feedback on 1/estimated quality of the automatic indexing proposed and 2/time gain achieved thanks to the indexing system.
- *Comparative evaluation*. Part of the French indexing systems benchmark evaluation corpus is in fact a French/English parallel corpus. This will make it possible to compare the French MeSH indexing systems to their English counterparts MeSHMap (Ruch et al., 2003) and MTI (Aronson et al., 2004).

A user interface will also be developed, in order to allow system customization for the indexers' use.

6. Conclusion

We have presented the architecture of an experimental MeSH automatic indexing system, using the INTEX platform, and MeSH and CISMeF terminologies. The system is intended as an indexing help for the CISMeF indexers, and focuses on the extraction of keyword/qualifier pairs, as opposed to single terms. The system achieves a precision that is comparable to that of other existing operational systems, and a recent evaluation showed that there is a major silence of the manual indexing on check tags. The CISMeF automatic indexing system will be used to enrich the current catalogue indexing with the missing check tags. However, several improvements, including significant MeSH dictionary coverage increase have to be performed before this tool can be considered as practically effective.

Acknowledgement

The authors would like to thank the indexers of the CISMeF team for their patience in detailing the manual indexing procedure, B. Dahamna for his technical support during the development of the automatic system, M. Silberztein for his help in the use of INTEX and K. Leuraud for her help in the statistical analysis of the data.

References

Abdallah, I. (1998). Segmentation et codage de signaux de parole par critères entropiques. PhD thesis, Université du Maine.

Anderson, J. D., & Pérez-Carballo, J. (2001). The nature of indexing: how humans and machines analyze messages and texts for retrieval. Part I: Research, and the nature of human indexing. *Information Processing & Management, 37*(2), 231–254.

Aronson, A. R., Mork, J. G., Gay, C. W., Humphrey, S. M., & Rogers, W. J. (2004). The NLM indexing initiative's medical text indexer. In *Proceedings of medinfo 2004* (pp. 268–272).

Courtois, B., & Silberztein, M. (1990). *Dictionnaires électroniques du français*. Paris: Larousse.

Crochemore, M., & Rytter, W. (1994). *Text algorithms*. New York: Oxford University Press.

Dailland, F., Leuthereau, A., & Vallée, H. (2003). Aide mémoire d'indexation MeSH et FMeSH pour le catalogage. Paris XI Medical School Library and INSERM Technical Report.

Darmoni, S. J., Leroy, J. P., Thirion, B., Baudic, F., Douyère, M., & Piot, J. (2000). CISMeF: a structured health resource guide. *Methods of Informative Medicine, 39*(1), 30–35.

Funk, M. E., Reid, C. A., & McGoogan, L. S. (1983). Indexing consistency in MEDLINE. *Bulletin of Medical Library Association, 71*(2), 176–183.

Gaudinat, A., & Boyer, C. (2002). Automatic extraction of MeSH terms from medline abstracts. NLPBA2002. In *Workshop on natural language processing in biomedical applications*.

Humphrey, S. M., & Miller, N. E. (1987). Knowledge-based indexing of the medical literature: The indexing aid project. *Journal of American Society of Information Science, 38*(3), 184–196.

Jacquemin, C., & Royauté, J. (1994). Retrieving terms and their variants in a lexicalised unification-based framework. In *Proceedings of the 17th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 132–141).

Koch, T. (2000). Quality-controlled subject gateways: definitions, typologies, empirical overview. *Online Information Review, 24*(1), 24–34.

Lahtinen, T. (2000). Automatic Indexing: an approach using an index term corpus and combining linguistic and statistical methods. PhD thesis, University of Helsinki.

Lancaster, F. W. (1998). *Indexing and abstracting in theory and practice*.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing* (pp. 534–536). Cambridge, MA: MIT Press.

Milstead, J. L. (1992). Methodologies for subject analysis in bibliographic databases. *Information Processing & Management, 28*(3), 407–431.

Névéol, A., & Ozdowska, S. (2005). Extraction de termes médicaux à partir d'un corpus parallèle anglais/français. In *Proceedings of the 5th conference on extraction et Gestion des Connaissances* (pp. 655–666).

Névéol, A., Rogozan, A., & Darmoni, S. J. (2004). Automatic Indexing of health resources in French for the CISMeF catalogue: a preliminary study. In *Proceedings of medinfo 2004* (p. 1772).

Névéol, A., Rogozan, A., Douyère, M., & Darmoni, S. J. (in press). Construction de ressources terminologiques en santé pour un système d'indexation automatique. In *Proceedings of the 7th INTEX workshop*.

Pouliquen, B. (2002). Indexation de textes médicaux par indexation de concepts, et ses utilisations. PhD thesis, Université Rennes 1.

Ruch, P., Baud, R., & Geissbühler, A. (2003). Learning-free text categorization. In M. Dojat, E. Keravnou, & P. Barahona (Eds.), *LNAI 2780* (pp. 199–204). Berlin: Springer.

Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Silberztein, M. (1993). *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson.

Soergel, D. (1985). *Organizing information: Principles of data base and retrieval systems*. Orlando: Academic Press.

Soualmia, L. F., Barry, C., & Darmoni, S. J. (2003). Knowledge-based query expansion over a medical terminology oriented ontology on the web. In M. Dojat, E. Keravnou, & P. Barahona (Eds.), *LNAI 2780* (pp. 209–213). Springer.

Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarousse, E., & Grabar, N. et al. (2003). Towards a unified medical lexicon for French. In *Proceedings of medical informatics Europe* (pp. 415–420).