Contribution to an Automated Indexing of French-language Health Web Sites

Michel Joubert, PhD,[1] Anne-Laure Peretti, MSc,[1] Stefan Darmoni, MD, PhD,[2]
Badisse Dahamna, MSc,[2] Marius Fieschi, MD, PhD,[1]

[1] LERTIM, Faculté de Médecine, Université de la Méditerranée, Marseille, France
[2] CISMeF, Centre Hospitalier Universitaire, Rouen, France

Objectives: *to improve the indexing of French-language health web sites by emphasizing the major terms that best describe them.*
Material and methods: *this study exploits both UMLS knowledge sources and results of previous research. It proposes a method for ranking MeSH terms taken from each record in order of relevance. The method is tested on a corpus of records taken from the French-language health gateway CISMeF.*
Results: *the results of the experiment are compared to those of a preliminary study performed on a corpus taken from MEDLINE.*
Discussion: *the ultimate objective of this work is to interface the developed tools with an automated MeSH term extractor in order to propose an automated indexing engine for French-language health web sites.*

INTRODUCTION

The main objective of the VUMeF[1] project is to improve the involvement of French in the Unified Medical Language System (UMLS) of the U.S. National Library of Medicine (NLM) [1]. This entails improving the French translation of the Medical Subject Headings thesaurus (MeSH) enriched with thousands of synonyms [2]. It also means the addition of a French translation of SNOMED, the validation of which is under way. This project follows a previous one, the UMLF [3], the aim of which was to produce a unified French-language medical lexicon. The automated indexing of web sites is an acknowledged task within the VUMeF project. This task is broken down into two sub-tasks: 1) the automated extraction of MeSH terms from documents published over the French-language web, and 2) the improvement of the operated indexing by emphasizing the *major* terms which best describe the content of each document as human librarians at the NLM do to summarize the main focus of each article. The automated French-language MeSH term extractor has been completed, its evaluation is currently in progress and its evaluation is currently underway on a parallel corpus with the NLM Medical Text Indexer [4].

This paper describes our efforts to achieve the objectives of the second above-mentioned sub-task. The method we propose is based on UMLS knowledge sources and exploits a knowledge source we developed in previous research works. A preliminary study allowed us to test the method we propose for ranking MeSH terms on a corpus of records taken from MEDLINE [5]. The results showed a favorable ranking of major terms. This paper presents the experiment we conducted using a corpus of records referencing humanly-indexed French-language web sites and accessible through the CISMeF[2] gateway [6]. This latter references French-language web sites according to given types: clinical guidelines, technical reports, educational resources, etc. The CISMeF indexers operate in similar fashion to MEDLINE indexers by indicating major terms associated with each document or web site. CISMeF is a manually-maintained and quality-controlled health gateway cataloguing the most important sources of health information in French in order to allow end-users to search them quickly and precisely. CISMeF indexes a great variety of resources (about 15,500) but has three main topics: guidelines for health professionals (about 2,400), teaching material for students (about 3,200), and consumer health information (about 2,500). A resource can be any medium containing health information e.g. a website, web pages, documents, reports and teaching material.

MATERIAL AND METHODS

We use several knowledge sources included in UMLS [7]: the semantic network (SN) and a database of co-occurrences of major terms in the MEDLINE literature. SN gives the binary relationships between types of concepts to which medical concepts from the Metathesaurus are attached. The Metathesaurus integrates numerous nomenclatures found in the biomedical domain, using

[1] Project sponsored by the French National Research Agency. http://www.vidal.fr/vumef/

[2] Catalogue et Index des Sites Médicaux Francophones. http://www.chu-rouen.fr/cismef/

a single identifier to reference concepts. The database of co-occurrences contains the frequencies of co-occurrences of pairs of <*term*/*subheading*>. We will name this database COOC in what follows. In previous studies, we built a table which translated UMLS semantic relationships into MeSH subheadings [8]. We will name this table SUBHEAD in what follows. For instance, the semantic relationship *Diagnoses* links the types of concepts: *Diagnostic Procedure* and *Disease or Syndrome*. This relationship is translated into the following subheadings that can be associated with relevant terms from the first type: *Pathology, Diagnosis, Radiography, Radionucleide imaging, Ultrasonography, Immunology, Microbiology, Virology*.

The aim of the method is to calculate a score for each keyword used to index a web site. The purpose is to rank the keywords according to their estimated significance in the description of the text from which they are extracted. When considering two MeSH keywords, and thus two concepts in the Metathesaurus, the following operations are performed:

- Identification of the SN concepts types they are attached to,
- Identification of the SN semantic relationships that link the identified concepts types,
- Translation of the identified relationships into MeSH subheadings for each term/concept using SUBHEAD,
- Selection of related lines in COOC for these terms and subheadings.

This sequence of operations is performed for each pair of MeSH keywords. A final operation adds the frequencies retrieved from COOC for a given keyword and a given relationship. This is done for each term and each related relationship. This provides a score which is calculated for each relationship that links terms. Finally, we assign a score to a keyword as follows: 1) we build an association graph the nodes of which are the keywords, and the edges are the semantic relationships with their associated scores; 2) the score of a keyword is the sum of the scores attributed to the edges that link it to other keywords.

To check whether the major terms assigned by human indexers have been properly ranked among the leading ones, after they have been scored as described above, we apply the hyper-geometric probability law, used for opinion polls and by librarians [9]. It can be described as the probability of m successes of k elements drawn from a pool of s favourable elements from a total of n elements. In our case, n is the number of keywords in each record, s is the number of major terms. If we expect to retrieve all the major terms in the list of the k first ones, then we postulate that we retrieve s major terms in the k leading ones (m=s). Thus, the formula is:

$$P(s \text{ of } k \text{ from } n) = C(n-s, k-s) / C(n, k)$$

Note that we presume our method to be infallible since with m=s we suppose that all the major terms are retrieved in the k first ranked terms. Hence, with a probability of 5% for a successful result due to chance, numbers n and s being part of each record, it is then possible to compute the number k of terms to be taken into account in the ranking: it is the lowest integer that satisfies the above formula. The *accuracy rate* P of the method applied to a record is P=q/s, where q is the number of major terms retrieved within the k first ranked terms, and s is the total number of major terms.

RESULTS

In a preliminary study we tested this method with a corpus of 1,444 records extracted from MEDLINE irrespective of origins and disciplines [5]. During this work we measured only the proportion of m major terms retrieved among the s leading ranked keywords. This means that k=s, s being the number of major terms in a record. We present here the results of the experiment we performed on the records of 404 technical reports and 1,012 educational resources taken from CISMeF after we had discarded the records containing only one keyword and the records containing keywords that were not associated in COOC.

Table 1 shows the number of records according to the number of keywords per record. Note that 76% of the records contain no more than 15 keywords. Table 2 shows the number of records according to the number of major terms per record. Note that 88% of the records contain no more than 5 major terms. Table 3 shows the distribution of the average number of keywords per record according to the number of major terms among them. The above proportions show that about 90% of the records taken from the CISMeF corpus are indexed by no more than 15 keywords, among which there are no more than 5 major terms, and that the number of major terms grows with the total number of keywords. We focus the analysis of the results on this set of records.

Figure 1 shows the average accuracy rate according to the total number of keywords per record, calculated as described above in accordance with the hyper-geometric law of probability. The curve shows that the average accuracy rate decreases as the number of keywords increases. Its average value is 0.69 ± 0.16. Figure 1 shows the average proportion according to the total number of keywords per record, by comparison with those retrieved from the CISMeF records. The curve shows that the average proportion also decreases as the number of keywords increases. Its average value is 0.65 ± 0.14.

Figure 2 shows the average accuracy rate according to the number of major terms per record of the CISMeF corpus. In contrast with the above result, this curve shows that the average accuracy rate is stable according to the number of major terms. It should be borne in mind that the related records constitute about 90% of the corpus. The average value is 0.65 ± 0.02. Figure 2 also shows the average proportion of retrieved major terms according to the number of major terms per record of the MEDLINE corpus. Its average value is 0.59 ± 0.08.

DISCUSSION

We successfully laid the foundations of the method in the framework of the European WRAPIN[3] project [10]. During this project we designed and developed tools to exploit UMLS knowledge sources in order to facilitate the indexing of health web sites and documentary databases [11]. Following the *Indexing Initiative* [12, 13], and complementary to the lexical analysis of free texts, the approach we adopted in WRAPIN attempted to exploit knowledge bases in order to improve indexing of documents. The MetaMap software extracts MeSH terms from health documents on the basis of a *ranking function* resulting of a *frequency factor* and a *relevance factor*. This latter is a weighted average of a MeSH tree depth factor, and other lexical features (word length, character count, …). Its authors planned to investigate the *semantic distance* between a given pair of UMLS concepts in order to quantify the notion of semantic locality. Our approach is comparable and draws not only on the MeSH thesaurus but also on the UMLS knowledge sources. This allows us to exploit both lexical aspects and semantic features concurrently.

Applied to both MEDLINE and CISMeF, with their different types of content, and based on different but similar measures, the results of our experiments are equivalent. Nevertheless, we expected a greater increase in accuracy during the current experiment. In our opinion, this shortfall is due to two facts: 1) the librarians at CISMeF do not follow the same indexing rules as those at MEDLINE, and 2) the knowledge database of co-occurrences is built on the MEDLINE literature and applied to the treatment of records taken from CISMeF.

Number of keywords	Number of records	Added percentages
2	14	0.01
3	66	0.06
4	80	0.11
5	86	0.17
6	116	0.26
7	132	0.35
8	101	0.42
9	107	0.50
10	96	0.56
11	70	0.61
12	70	0.66
13	49	0.70
14	54	0.74
15	43	0.76

Table 1. Distribution of number of records according to the number of keywords per record.

Number of major terms	Number of records	Added percentages
1	398	0.28
2	383	0.55
3	244	0.72
4	128	0.81
5	89	0,88

Table 2. Distribution of number of records according to the number of major terms per record.

Number of major terms	Average number of keywords
1	9.12
2	9.34
3	11.37
4	13.30
5	16.29

Table 3. Average number of keywords according to the number of major terms per record.

[3] Worldwide Reliable Advice for Patients and Individuals. European project IST-2001-33260. http://www.wrapin.org/

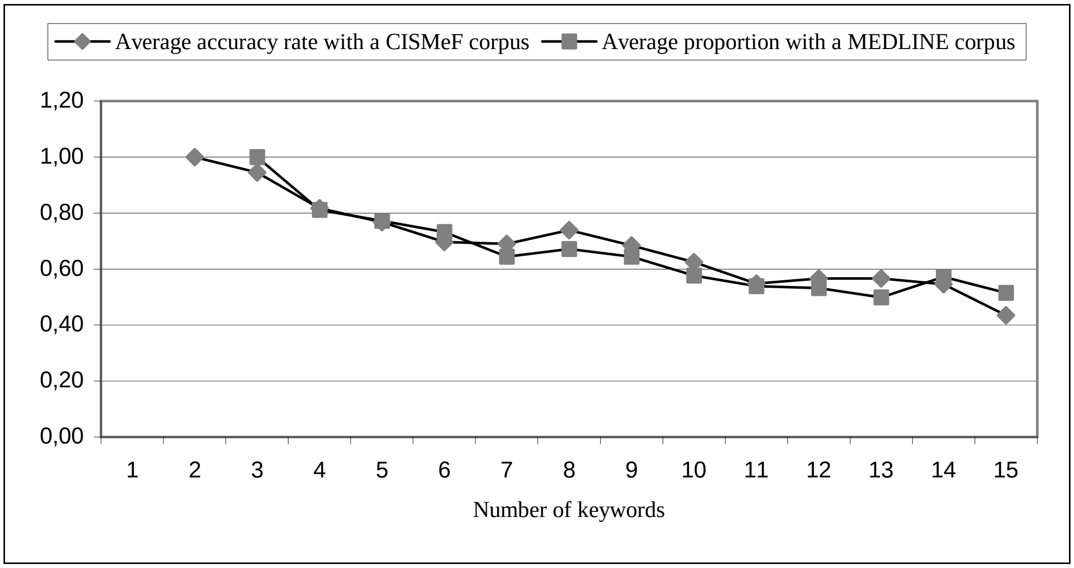

Figure 1. Average accuracy rate and proportion according to the number of keywords per record.

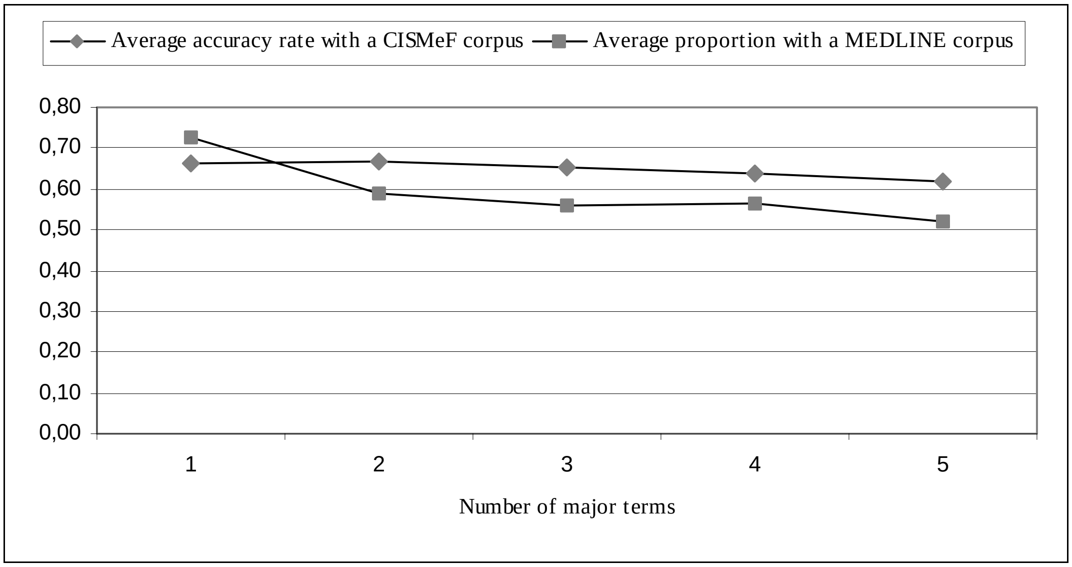

Figure 2. Average accuracy rate and proportion according to the number of major terms per record.

This leads us to believe that different indexing rules and different types of documents for indexation have a real impact on the results our method produces. The medical librarians at CISMeF do not follow the same indexing rules as those at MEDLINE, because they have to face a much larger heterogeneousness of health resources to index on the basis of CISMeF types. This is in contrast with the practice of librarians at MEDLINE who index only scientific articles. The CISMeF editorial policy varies according to the type of resource: e.g. the CISMeF team will index in more detail clinical guidelines than web sites of patient associations. The mean of MeSH terms is 16.56 ± 20.57 per clinical guidelines versus 2.51 ± 2.26 for patient associations.

On the whole, our method is based on two knowledge sources: 1) the UMLS Semantic Network which proposes relationships, and 2) a co-occurrences database which filters proposed semantic relationships according to a documentary collection from which it has been built. The shortfall we observe in our experiments between the MEDLINE and CISMeF corpuses is probably due to the fact that the co-occurrences database was built on the MEDLINE literature and exploited with records from the CISMeF corpus with which it is not in accordance. A possible way to overcome this difficulty is to build and exploit a co-occurrences database using major terms in the CISMeF database, and then compare the obtained results with the current ones. This approach is in accordance with the definition of Gruber who wrote that an ontology provides a representational vocabulary for a given domain with a set of rules that constrain the meaning of the terms in that vocabulary sufficiently to enable consistent interpretation of data framed in that vocabulary [14]. As static representation of concepts is supplied by the Metathesaurus and the SN, the rules of use of semantic relationships are given by co-occurrences between concepts.

A next step is to interface the tools we developed with the MeSH term extractor the validation of which is underway. The aim is to produce an automated indexing engine of French-language health web sites. The extended capabilities and the specialization to the health domain of such a search engine would then propose answers ranked in order of relevance as responses to queries to a gateway of health web sites, based on medical semantics in contrast with lexical and structural features that current search engines exploit.

ACKNOWLEDGEMENTS

The authors thank the whole consortium of the VUMeF project. They also thank the US National Library of Medicine which kindly provided them with the UMLS knowledge sources.

REFERENCES

[1] Darmoni SJ, Jarrousse E, Zweigenbaum P, Le Beux P, Namer F, Baud R, Joubert M, Vallee H, Cote RA, Buemi A, Bourigault D, Recource G, Jeanneau S, Rodrigues JM. VUMeF: extending the French involvement in the UMLS Metathesaurus. *Proc. AMIA Annu Symp*. 2003: 824.

[2] INSERM. Le MeSH bilingue français-anglais. http://ist.inserm.fr/basismesh/mesh.html.

[3] Zweigenbaum P, Baud R, Burgun A, Namer F, Jarrousse E, Grabar N, Ruch P, Le Duff F, Forget JF, Douyere M, Darmoni S.UMLF: a unified medical lexicon for French. *Int J Med Inform* 2005; 74(2-4): 119-24.

[4] Névéol A, Mork JG, Aronson AR, Darmoni SJ. Evaluation of French and English MeSH Indexing Systems with a Parallel Corpus. *Proc. AMIA Annu Symp*. 2005: 565-9.

[5] Joubert M, Peretti AL, Gouvernet J, Fieschi M. Refinement of an automatic Method for Indexing medical Literature – a preliminary Study. *Proc. MIE 2005*: 683-8.

[6] Darmoni SJ, Leroy JP, Baudic F, Douyere M, Piot J, Thirion B. CISMeF: a structured health resource guide. *Meth Inform Med*. 2000; 39: 30-5.

[7] McCray AT, Nelson SJ. The Representation of Meaning in the UMLS. *Meth Inform Med* 1995; 34: 193-201.

[8] Aymard S, Fieschi D, Volot F, Joubert M, Fieschi M. Towards interoperability of information sources within a hospital intranet. *Proc. AMIA Annu Symp*. 1998: 638-42.

[9] Wilbur WJ. Retrieval testing with hyper-geometric document models. *Journal of the American Society for Information Science*, 44(6); 1993: 340-51.

[10] Gaudinat A, Ruch P, Joubert M, Uziel P, Strauss A, Thonet M, Baud R, Spahni S, Weber P, Bonal J, Boyer C, Fieschi M, Geissbuhler A. Health search engine with e-document analysis for reliable search results. *Int J Med Inform* 2006; 75(1): 73-85.

[11] Gaudinat A, Joubert M, Aymard S, Falco L, Boyer C, Fieschi M. WRAPIN: new health search engine generation using UMLS Knowledge Sources for MeSH term extraction from health documentation. *Proc. Medinfo 2004*. IOS Press; 2004: 356-60.

[12] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM Indexing Initiative. *Proc AMIA Annu Symp*. 2000: 17-21.

[13] Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. *Proc. Medinfo*. 2004: 268-72.

[14] Gruber TR. A Translation Approach to Portable Ontologies. *Knowledge Acquisition* 1993; 5: 199-220.

Address for correspondence

Michel JOUBERT
LERTIM
Faculté de Médecine
27, bld Jean Moulin
13005 Marseille
FRANCE
e-mail: mjoubert@ap-hm.fr
http://Cybertim.timone.univ-mrs.fr/