Multiple Terminologies in a Health Portal: Automatic Indexing and Information Retrieval

Stéfan J. Darmoni, MD, PhD[1], Suzanne Pereira, PhD[1,2,3], Saoussen Sakji, MSc[1], Tayeb Merabti, MSc[1], É. Prieur, PhD[1], Michel Joubert, PhD[2], and B. Thirion, Head Librarian[1]

[1] CISMeF, LITIS EA 4108, University of Rouen, Normandy, France
[2] LERTIM, Marseille Medical University, France
[3] VIDAL, Issy les Moulineaux, France

Abstract. Background: In the specific context of developing quality-controlled health gateways, several standards must be respected (e.g. Dublin Core for metadata element set; thesaurus MeSH as the controlled vocabulary to index Internet resources; HON code to accredit quality of health Web sites). These standards were applied to create the CISMeF Web site (French acronym for Catalog & Index of Health Internet resources in French). Objective: In this work, the strategic shift of the CISMeF team is intended to index and retrieve French resources not anymore with a single terminology (MeSH thesaurus) but with the main health terminologies available in French (ICD 10, SNOMED International, CCAM, ATC). Methods & Results: Since 2005, we have developed the French Multi-Terminology Indexer (F-MTI), using a multi-terminology approach and mappings between health terminologies. This tool is used for automatic indexing and information retrieval. Conclusion: Since the last quarter of 2008, F-MTI is daily used in the CISMeF production environment and is connected to a French Health Multi-Terminology Server.

1 Introduction

Regardless of their web experience and general information retrieval skills, users have difficulties in seeking health information on the Internet [1]. In this context, several quality-controlled health gateways have been developed [2]. Quality-controlled subject gateways (or portals) were defined by Koch [2] as Internet services which apply a comprehensive set of quality measures to support systematic resource discovery. Among several quality-controlled health gateways, CISMeF ([French] acronym for Catalog and Index of French Language Health Resources on the Internet) was designed to catalog and index the most important and quality-controlled sources of institutional health information in French in order to help health professionals, patients and students to find electronic medical information available on the Internet quickly and precisely [3]. To respect quality standards, CISMeF is accredited by the Health on the Net Foundation since

1998 [4]. In the catalog, all the resources are described with 11[1] out of 15 Dublin Core (DC) metadata set [5] including the title and resource types, and indexed with a set of indexing terms to describe the information content. Since 1995 (creation of CISMeF in February), the indexing terms are descriptor/qualifier pairs or descriptors from the MeSH® thesaurus (Medical Subject Headings), the U.S. National Library of Medicine's (NLM's) controlled vocabulary used to index articles from the biomedical literature. From 1995 to 2002, CISMeF was exclusively manually indexed by a team of four indexers, which are medical librarians and systematically checked by the chief information scientist[2]. The objective of this paper is to describe the strategic shift to use several health terminologies for the automatic indexing and the information retrieval in the CISMeF quality-controlled health portal vs. the previous use of only one medical terminology (the MeSH thesaurus).

2 Methods

2.1 Automatic Indexing

Since 2002, faced with the growing amount of online resources to be indexed and included in the catalog, the CISMeF team consistently evaluated advanced automatic MeSH indexing techniques. The automatic indexing tools used primarily natural language processing (NLP) and K-nearest neighbours (KNN) methods [6], followed by a simpler bag of words algorithm [7]. The latter was successfully evaluated in the context of teaching resources. In August 2006, the CISMeF team decided to use this algorithm in the daily practice for most of the Internet resources rated as "low priority" resources (except guidelines which are still manually indexed because this type of resources rated as "high priority" need in-depth indexing). These "low priority" resources are teaching resources or resources belonging to a topic substantively covered in the catalog that do not require in-depth indexing.

Since 2005, the CISMeF team and the Vidal company developed the F-MTI tool [8]; the goal of the CISMeF team was to use a new automated indexing tool to index health resources in CISMeF; from a bag-of words algorithm based on a mono-terminology approach, we choose to use the F-MTI tool based on several health terminologies. In 2006, besides the MeSH, four health terminologies were included in F-MTI: ICD-10 (International Classification of Diseases) and SNOMED 3.5 (Systematized Nomenclature of Medicine) which are included in the UMLS, CCAM (the French equivalent of US CPT) and TUV (a French terminology for therapeutic and clinical notions for the use of drugs), which are not included in the UMLS. These four terminologies are mapped to the French MeSH. Several formal evaluations were performed with each of these terminologies [8], including the latest with the MeSH thesaurus [9]. During 2008, four

[1] Inclusion: title, creator, subject, description, publisher, date, type, format, identifier, source and language. Exclusion: relation, coverage, rights and contributor.

[2] Its URLs are http://www.chu-rouen.fr/cismef or http://www.cismef.org

new terminologies or classifications were added: ATC classification (N=5,514), drug names with international non-proprietary names (INN) and brand names (N=22,662), Orphanet thesaurus for rare diseases (N=7,421), MeSH Supplementary Concepts translated in French by the CISMeF team (N=6,004 out of over 180,000). Currently, after the formal evaluation of the F-MTI to index health resource [9], this tool F-MTI has enabled the automatic indexing of 33,951 resources in the CISMeF catalog and the semiautomatic (or supervised) indexing[3] of another 12,440 resources based on resources titles; overall, 65,242 resources are included in this gateway. Three levels of indexing were defined in the CISMeF catalogue:

- Level 1 or Core-CISMeF (N=18,851) which is totally manually indexed resources (e.g. guidelines).
- Level 2 or supervised resources (N=12,440): these resources are rated by the CISMeF editorial board as less important than level 1. These resources do not need in-depth indexing (e.g. technical reports, teaching resources designed at the national level, document for patients from medical specialties).
- Level 3 or automatically indexed resources (N=33,951). The CISMeF editorial board has rated these resources as less important than level 1 and level 2 (e.g. teaching resources designed at the medical school level, patient association Web sites).

Since CISMeF achieved this milestone in automatic indexing, it strived to improve the automatic indexing algorithm and make it on par with manual indexing. One of the challenges that the CISMeF automatic indexing algorithm needs to address is identifying all the different forms a term can take in natural language, specifically with respect to lexical and grammatical variations. Most terminologies such as MeSH provide synonyms and variants for the terms but this information is usually insufficient to describe all the forms that can be encountered for a given term in a document.

To allow automatic indexing using multiple terminologies to be used in the CISMeF catalog, the CISMeF team in collaboration with eight 4th year students of the INSA of Rouen engineering school has to integrate the previously listed health terminologies in the CISMeF back-office. To do so, for each terminology, a UML model and a parser were developed. A generic model was also developed to allow inter-terminology interoperability. Each specific terminology is integrated in an OWL format into a health multi-terminology server (French acronym SMTS) using the ITM® model (Mondeca) for implementation. From this SMTS, all the health terminologies were uploaded in the CISMeF backoffice. Then, F-MTI automatic indexing results are able to be easily integrated in the CISMeF backoffice and allowing multi-terminology information retrieval.

[3] supervision means that these resources are primarily indexed automatically, and then this indexing is reviewed by a CISMeF human indexer, who is a medical librarian.

2.2 Information Retrieval

Since the overall CISMeF structure has evolved from a mono-terminological world (based on the MeSH thesaurus) to a multi-terminological universe, similarly the information retrieval algorithm has been modified. The formulation of the requests consists in re-writing a user query in order to conceive queries closer to the expected needs. The Doc'CISMeF search engine carries out a comparison between character strings. Currently, the task of query reformulation is carried out by listing all the possible combinations of the bag of words query terms (terms obtained following the initial user query treatment by eliminating the blank words and by stemming the terms) in order to find the maximum of possible correspondences with the documents descriptors- the terms considered as the most significant. Indeed, pairing resource-query is performed by a comparison of what could exist as correspondence between the query terms and the resources descriptors. Several operations are done during this process such as: natural language treatment techniques for the multi terms query, phonemisation, terms adjacency. To match as much as possible queries with the CISMeF corpus, we have implemented a three-step heuristics. The process consists in recognizing a user query expression.

- Step 1. The reserved terms or the document's title: If the user query expression matches CISMeF terminology terms or the document's title, the process stops, and the answer to the query is the union of the resources that are indexed by the query terms, and those that are indexed by the terms they subsume, directly or indirectly, in all the hierarchies they belong to. This step is modified since the implementation of the multi-terminology. This process is generalized to match to any descriptor belonging to multiple health terminology, and the generalization includes also the list of terms.
- Step 2. The CISMeF metadata: The search is performed over all the other fields of the CISMeF metadata (abstract, author, publisher, identifier ...).
- Step 3. Adjacency of words in the text: A full text search over the document with adjacency of n words with n= 10 $\times$ (number of words of the query - 1) is realised.

The steps 2 & 3 are similar in the multi-terminology context than previously in the mono-terminology context. By default and for each of the three steps, CISMeF displays the query results starting with the most recent document and displays the resources indexed with the MeSH Major headings, which express the main topic and then the resources indexed with the MeSH minor headings, which produce a complementary information about the indexing. The resources that were indexed automatically in the catalogue are displayed after those that were indexed manually.

3 Conclusion and Perspectives

As far as we know, the F-MTI automatic indexing tool is the first attempt to use multiple health terminologies besides the English (specially the MTI initiative of

the US National Library of Medicine). In the near future, a formal evaluation of F-MTI will be performed on French scientific articles included in the MEDLINE database: a French/English comparison is feasible on this MEDLINE subset.

Information retrieval using multiple terminologies will also be compared to the previous information retrieval based only on the MeSH thesaurus. The information retrieval using multiple terminologies will be adapted on a new context: to retrieve reports from the electronic health record of patients.

4 Acknowledgments

This research was partially supported by the European Union funded project FP7-ICT-1-5.2-Risk Assessment and Patient Safety (PSPIP) (n°216-130), and the ANR-funded project ANR-07-TECSAN-010. The authors would like to thank CISMeF indexers for their help in the study design and result analysis.

References

[1] Keselman A., Browne A.C., Kaufman D.R.: Consumer Health Information Seeking as Hypothesis Testing. J Am Med Inform Assoc.vol. 15, no. 4, pp. 484-495, Jul-Aug 2008. Epub 2008 Apr 24.

[2] Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy JP, Thirion B, Darmoni SJ: Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. Health Info Libr J 2004 Dec: 21(4):253-61.

[3] Koch T: Quality-controlled subject gateways: definitions, typologies, empirical overview, Subject gateways. Online Information Review 2000: 24(1): 24-34.

[4] Boyer C, Gaudinat A, Baujard V, Geissbühler A.: Health on the Net Foundation: assessing the quality of health web pages all over the world. Stud Health Technol Inform. 2007;129(Pt 2):1017-21.

[5] Dekkers M, Weibel S.: State of the Dublin Core Metadata Initiative. D-Lib Magazine 2003;9(40).

[6] Névéol A, Rogozan A, Darmoni SJ.: Automatic indexing of online health resources for a French quality controlled gateway. Information Management & Processing 2006:1; 695-709.

[7] Névéol A, Pereira S, Kerdelhué G, Dahamna B, Joubert M , Darmoni SJ.: Evaluation of a simple method for the automatic assignment of MeSH descriptors to health resources in a French online catalogue. Medinfo 2007, 407-411.

[8] Pereira S.: Multi-terminology indexing of concepts in health. [Indexation multi-terminologique de concepts en santé]. PhD Thesis, University of Rouen, Normandy, France

[9] Pereira S, Neveol A, Kerdelhué G, Serrot E, Joubert M, Darmoni SJ.: Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. AMIA Annu Symp Proc. 2008 Nov 6:586-90.

[10] Joubert M, Dahamna B, Delahousse B, Fieschi M, Darmoni SJ.: SMTS®: Un Serveur Multi-Terminologies de Santé. In: Informatique & Santé, Journées Francophones d'Informatique Médicales, (in press).

[11] Soualmia L, Dahamna B, Thirion B, Darmoni SJ.: Strategies for health information retrieval. Stud Health Technol Inform, Volume 124, Pages 595-600, 2006.