Multilingual Enrichment of an Ontology of Cardio-Vascular Diseases

M Simonet[1], R Patriarche[1], A Bechlioulis[2], D Bernhard[1], G Diallo[1], R Messai[1], S Gedzelman[1]

[1]Université Joseph Fourier, Grenoble, France
[2]Michailideion Cardiac Center, Ioannina, Greece

Abstract

Concept-based Information Retrieval offers a higher abstraction level than the classical keyword-based approach. A multilingual terminology makes possible language-independent indexing and querying. The conceptual structure with its associated terminology is called an ontology. This paper describes a methodology to enrich an existing ontology represented in OWL with terms in different languages, and its application to a MeSH-based ontology of the cardio-vascular domain. The first enrichment phase is automatic and makes use of the UMLS multilingual lexicon. The second phase is partially manual and relies on a corpus of texts. In the Noesis European project, where this work was carried on, the multilingual ontology was used for automatic indexing and manual semantic annotation of cardio-vascular documents.

1. Introduction

Information Retrieval (IR) aims at returning documents satisfying a user's query. In classical Information Retrieval, documents are taken from a repository of candidate documents built beforehand. A typical example is the Pubmed repository of biomedical scientific texts accessed through Medline. Prior to user's search, the documents have been indexed. Indexing is the process of associating with each document keywords taken from a fixed set of terms. The set of keywords associated with a document constitutes an "image" of the document, which will be used to answer a user's query. To be effective, a query must use the same terms as those used for indexing. Pubmed texts are indexed by keywords which represent entries of the MeSH thesaurus and the indexing process is done manually (about 70 indexers are working permanently to achieve this task).

The MeSH thesaurus has also been used for automatic indexing. CiSMeF uses a French version of the MeSH [http://ist.inserm.fr/basismesh/mesh.html] to manually index the French biomedical scientific literature and is currently working on algorithms for automatic indexing based on the MeSH terms and structure [2], again using the French associated vocabulary. Health on the Net (HON) also performs MeSH-based automatic indexing of the biomedical literature and web sites [3].

Although the MeSH thesaurus remains the reference classification to index the biomedical literature, the need to establish links between existing medical classifications has led to the UMLS initiative. Since the beginning, the UMLS was designed to associate a multilingual vocabulary with the concepts of its so-called meta-thesaurus, which constitutes its very heart (1 700 000 concepts in the 2006 version) **Error! Reference source not found.**. The UMLS multilingual lexicon was used in our first enrichment phase to automatically enrich the MeSH vocabulary.

Concepts can be identified in a text (for indexing) or in a query only through the terms (words and groups of words) that represent them. In order to fully benefit from the possibilities of concept-based indexing and querying, the terminology associated with the concepts must be as rich as possible in order to match the actual terms in the documents and in the user's query. Performing indexing and querying, at the concept level, makes it possible to express a query in other languages, without sticking to a fixed set of keywords, and even by using everyday language [7].

Concept-based multilingual information retrieval was one of the objectives of the Noesis European project [http://www.noesis-eu.org]. The MeSH thesaurus was chosen as the basis for the conceptual structure to support indexing and querying. The initial work to represent the MeSH concepts and vocabulary in OWL (Ontology Web Language), which is the language recommended by the world wide web consortium (w3c) to represent a concept-based structure, was presented at CinC'05 [4], along with the first attempts for automatic multilingual enrichment through the UMLS. In this paper we present the methodology for multilingual enrichment, which has since been refined, and the software tools which have been designed both for the automatic phase (through the UMLS) and the semi-automatic phase, where they assist the ontologist (the person who does the manual part of the enrichment task) in his/her work.

2. Material and methods

There is no agreed upon definition of an ontology so far, and there are still discussions about the very basic constituents of an ontology. Klein and Smith tried to clarify the world of concept systems and ontologies and recently proposed definitions of the concepts used in both views [5]. As we only considered the enrichment of the terminology part, our work applies to both approaches. Moreover, as the skeleton of our concept structure is the MeSH generic/specific relationship between descriptors, we did not have to distinguish between *Is_A* and *Part_Of* relationships between the categories we used as concepts, although an in-depth ontological work should make such a distinction. The generic/specific relationship was represented by the class/subclass relationship in OWL[1], which was adequate to Information Retrieval and annotation purposes [8].

In a thesaurus, the basic elements are terms (i.e., words and expressions) while in an ontology they are concepts. The synonymy relationship that is explicit in a thesaurus implicitly holds between the terms associated with the same concept in an ontology. For example, the terms *blood platelets, platelets, thrombocytes* and *PLT* are considered synonyms, as they can be used in texts to refer to the concept BLOOD PLATELETS. Among those terms, one is chosen to represent the concept in user's interfaces: it is called the concept's preferred term (*blood platelets* in the above example). In a multilingual approach, a preferred term is required for each considered language. As OWL does not implement this feature, we used the SKOS (Schema for Knowledge Organization Systems) format [http://www.w3.org/2004/02/skos], designed to express the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies and other types of controlled vocabulary, in complement to the OWL representation.

The starting point of our concept structure and terminology was a strict subset of the cardiovascular MeSH descriptors [4], which provided 690 concepts, represented as 690 OWL classes. However, the first tests with users showed that some cardiovascular concepts were missing and that general biomedical concepts should also be added, as they were liable to be used in queries, whatever the specific domain considered. This led to 3 489 concepts out of 22 995 descriptors for the entire 2005 MeSH. 12 227 English terms were associated with these concepts, 6736 French terms could also be obtained from the bilingual English-French version of the MeSH. This was the basis for the multilingual enrichment that is presented below.

[1] OWL (Ontology Web Language) is the language recommended by the w3c (world wide web consortium) to represent ontologies for computer usage. Note that other representations, such as Sowa's Conceptual Graphs, are possible alternatives for ontology representation.

Enrichment steps

The enrichment process consisted of three steps:

1) based on the the UMLS multilingual lexicon,
2) based on a corpus of English texts,
3) through manual translation.

Step 1, based on the UMLS, concerned only the languages considered in the UMLS lexicon, which covered the languages of the Noesis consortium (English, Spanish, French, German and Italian) except Greek.

It consisted in adding to each concept the terms attached to the corresponding concept in the UMLS. This was possible because each concept in the MeSH thesaurus and the UMLS meta-thesaurus has a unique MeSH and UMLS identifier, and that the UMLS provides a link between them. In our concept structure, we kept both identifiers, so as to make possible further enrichment following the constant evolution of the MeSH and the UMLS.

However, the terms obtained from the UMLS lexicon could not be used as such in our terminology, which aims at supporting the indexing and querying IR processes, while the UMLS lexicon contains terms with extra information that make it difficult to use them for information retrieval. As an example, we present the UMLS terms associated with the concept AORTA:

- Aorta
- Aortas
- Aorta, NOS
- 42 AORTA
- Aorta, Ascending
- Aortas, ascending
- AORTA ASCENDING
- Ascending Aorta
- AA - Ascending aorta
- Aortic structure (body structure)
- Aortic (qualifier value)

from which we retained only:

- aorta
- ascending aorta
- aorta ascending
- aortic
- aortic structure

which are useful for information retrieval. To do so, we eliminated the terms containing one of the characters { (,), [,], <, > } or the string NOS (which means *Not Otherwise Specified*).

As Noesis stems[2] the terms in the texts before indexing and in the users' queries, we only kept one term among

[2] Stemming is the process of eliminating word endings.

those that were stem-equivalent, i.e., which provided the same stemmed form. Lexical variants of a term are a typical example of this situation, e.g., *aortas*, *aorta* and *aortic*. In this case, only the term *aorta* is retained.

We also noticed that the MeSH and the UMLS vocabulary contained forms of some terms that could not be found as such in texts, such as *aorta, ascending* (with a comma). These terms were obviously introduced as standard forms oriented towards manual indexing. As the textual form of a term, i.e., the form which is liable to be actually found in a text, is most often present among the list of alternative terms, we removed those terms containing a comma. However, before removal, we checked that there was a form without comma that contained the same set of stems, although in a different order.

This automatic cleaning of the English vocabulary eliminated 48 310 terms out of the 73 886 initially provided by the MeSH+UMLS for the 3 489 concepts, thus scaling down the vocabulary to "essential" terms in the IR perspective. The ratio of eliminated terms was less important for the other languages.

Step 2 aimed at finding new terms in texts of the considered domain A corpus is a collection of texts designed for some purpose. Ideally, it should be of the same kind as the documents to be queried later. The set of documents accessed through Pubmed would constitute an ideal corpus, but for copyright reasons only abstracts can be downloaded freely. We could find a set of freely accessible scientific documents related to the cardiovascular domain through the biomedcentral portal [http://www.biomedcentral.org].

To build the English corpus we selected the journals related to the Cardio-Vascular domain. This provided 490 documents, extracted from five journals:

- Cardiovascular Diabetology (45 titles)
- Cardiovascular Ultrasound (64 titles)
- Current Controlled Trials in Cardiovascular Medicine (156 titles)
- BioMed Central (200 titles)
- Thrombosis Journal (25 titles)

The term extraction algorithm provided 114 509 terms, by using the method of repeated segments. This method detects units composed of several words repeated in the same order in different locations within the corpus of texts under analysis. Some of these multi-word units occur frequently and are considered as candidate terms.

This method is language-independent and it has the reputation of being fairly robust. We provided it with two stop-lists: 1) words which should not be contained in a term, (e.g., *am, and, are, be, because, but, discussion, during, email, e-mail, however, if, …*), and 2) words which a term should not start or end with (e.g., *about, above, across, after, again, …*).

We first eliminated redundant terms, i.e., the terms whose stemmed form was already present in the ontology (via stemming). We also performed a semantic filtering by eliminating the terms of which none constituent words matched a word in at least one term in the ontology.

After these two phases of cleaning, there remained 82 489 terms that were manually examined, finally leaving 2 844 terms to be inserted in the ontology.

A software environment, the Noesis Enrichment Tool (NOET), was designed to support the human activity of term selection and assignment to a concept [8]. Some of its components proved very useful to the ontologist:

- a concordancer displays a term in its context, thus ensuring a right understanding of its meaning,
- a terminology server finds concepts close to a term, thus facilitating the choice of the right concept to assign to a term.

Step 3 is called *translation* although it is not a term-to-term translation. It consists in manually adding vocabulary for non-English languages, using the English terms as a basis.

3. Results

Table 1 shows the enrichment results of step 2 (based on the UMLS), for the 3489 concepts currently being considered. This number is liable to increase after the validation phase in the Noesis project, as users' queries might use terms referring to concepts absent of the ontology.

Table 1. Vocabulary enrichment (3489 concepts)

	Initial (MeSH)	Step 1 (UMLS)	Step 2 (corpus)
English	12 227	25576	+ 2 844
Spanish	0	14541	
German	0	10332	
French	6736	10065	
Greek	0	0	

As can be seen, all the languages are not equally treated in the UMLS, and Greek is totally absent. The original lexicon is in English and the enrichment in other languages depends on national initiatives. The French project UMLF will contribute to significantly increase the French vocabulary [10].

The second contribution of this work is a set of software tools that can be used to produce an OWL representation of the MeSH thesaurus and perform vocabulary enrichment. Multilingual aspects of the terminology associated with an ontology are not currently

properly supported by the Semantic Web main stream. The NOET (Noesis Enrichment Tool) offers such support, relying on the SKOS initiative.

An OWL+SKOS version of the whole MeSH will be produced soon, with the enrichment level corresponding to step 2 (based on UMLS). It will also be possible to use the NOET tool to support enrichment for other languages than those considered in the Noesis project.

4. Discussion and conclusions

In the design of the Noesis project, the Knowledge Management part relied on an ontology of cardiovascular diseases. Its purpose was to support concept-based information retrieval in order to make possible language-independent indexing and querying. As no such ontology was available, a subset of the MeSH thesaurus was used as the basis of the conceptual structure needed for this purpose and it was represented in the OWL standard ontology language. Although it is represented in OWL in order to benefit from the tools and methods developed in the context of the Semantic Web, this structure cannot be considered as an ontology [5] and it should rather be called a *concept-based terminology*.

The results of this work are twofold: the enriched terminology itself and the methods and tools that were designed for the enrichment task. As the translation phase is still going on, we could only evaluate the results of the first phase, through the UMLS (Table 1). The methods and tools that we have presented can be straightforwardly applied to the multilingual enrichment of any conceptual structure – including an ontology – provided that it uses an OWL+SKOS representation similar to the one we used. To benefit from the first enrichment phase through the UMLS, the concepts to be enriched must have a link to the equivalent UMLS entry.

In order to be used for automatic indexing, the whole MeSH thesaurus should be enriched, through the UMLS first and secondly through a corpus. The documents indexed through Pubmed could be used for that purpose, thus guaranteeing an almost exhaustive coverage of the vocabulary of the biomedical field. Such enrichment is not necessary for the current use of the MeSH thesaurus, where the human skills compensate the lack of vocabulary for the manual indexing task. This enrichment would also be beneficial in the perspective of information extraction [1], which needs a much more refined knowledge representation and terminology covering than classical Information Retrieval. However, for this usage, the MeSH conceptual hierarchy should also be improved, as it is not ontologically sound. The recent OBO foundry initiative offers such a perspective by aiming at building an ontology of the whole biomedical domain [9]. As an ontology, according to Barry Smith's view, should not deal with terminology, our terminology work, once extended to the whole MeSH – hence covering the whole medical domain – will be available as a complement to any ontological work in this field.

Acknowledgements

This research was funded by the European Commission - (NOESIS project, IST-2002-507960).

References

[1] Cunningham H. Automatic Information Extraction. Encyclopedia of Language and Linguistics, 2nd Edition, Elsevier, 2005.

[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 Information and Libraries Journal 2004 Dec;21(4):253-61.

[3] Gaudinat A, Joubert M, Aymard S, Falco L, Boyer C, Fieschi M. WRAPIN: new generation health search engine using UMLS knowledge sources for MeSH term extraction from health documentation. Medinfo. 2004;11(Pt 1):356-60. PMID: 15360834.

[4] Gedzelman S, Simonet M, Bernhard D, Diallo G, Palmer P. Building an ontology of Cardio-Vascular diseases for Concept-Based Information Retrieval, Computers in Cardiology, Lyon, France, Sept. 2005.

[5] Klein G, Smith B. Concept Systems and Ontologies. J Biomed Inform. 2006;39(3):274-87. 6

[6] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med. 1993 Aug;32(4):281-91.

[7] Messai R, Simonet M, Mousseau M. A breast cancer terminology for lay people. 5th European Breast Cancer Conference, Nice. 2006. European Journal of Cancer EJC Supplements. Vol (4):2, P:179-180.

[8] Simonet M, Patriarche R, Bernhard D, Diallo G, Ferriol S, Palmer P. Multilingual ontology enrichment for semantic annotation and retrieval of medical information. Mednet'06, Toronto, Canada, Oct. 2006.

[9] Smith B, Ceusters W. Ontology as the Core Discipline of Biomedical Informatics. Legacies of the Past and Recommendations for the Future Direction of Research. Forthcoming in Computing, Philosophy, and Cognitive Science, G.D. Crnkovic and S. Stuart (eds.), Cambridge: Cambridge Scholars Press, 2006.

[10] 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 Mar; 74(2-4):119-24.

Address for correspondence

Michel Simonet
TIMC laboratory – Faculté de Médecine
38700 La Tronche
France
Michel.Simonet@imag.fr