Detecting MeSH Keywords and Topics in the Context of Website Quality Assessment

Dušan Rak[1], Vojtěch Svátek[1], Manuel Fidalgo[2], Olli Alm[3]

[1] University of Economics, Prague (UEP), Czech Republic
[2] Universidad Nacional de Educación a Distancia (UNED), Spain
[3] Helsinki University of Technology (TKK), Finland

ABSTRACT: Automatic detection of keywords and general topics is a special-purpose auxiliary task in the website quality assessment process. We describe the approach to obtaining such information used in the MedIEQ project, discuss problems related to the type of human language used in medical websites, and illustrate them on examples.

1. Introduction

Identification of *keywords* and *topics* for a certain textual resource is a notorious task in computer science. The most frequent motivation of this task is to support document retrieval in large collections. By topic detection we mean the process that describes a document based on the keywords found from the document. An associated but not necessarily dependent task is that of document *categorisation*, typically into classes theoretically assumed to be disjoint.

In this paper, we rather consider *keyword* and *topic* identification in a specific application context, *website quality assessment*, and appearing side-by-side with more focused *information extraction* techniques. Aside the domain-neutral usage of keywords and topics for document retrieval and management purposes, such an application also considers them within the actual quality assessment workflow. The most obvious case is a comparison between the *declared* topics of the website content and the *actual* topics as identified bottom-up in the text. Clearly, a website that does not explicitly and correctly define its topical scope may not be trustworthy enough for the consumer, even if its formal characteristics (such as responsibility identification or presence of policy statements) are compliant with standards.

The obvious way to identify *keywords* in medical web pages is to employ some public and generally accepted controlled vocabulary or ontology such as the *Medical Subject Headings*[1] (MeSH). Although the matching process is not simple (e.g. due to polysemy and homonymy), it has been proven that results comparable to human-level indexing can thus be obtained.

Even a keyword list may be useful guidance for the certification authority in deciding whether the declared topic matches well with the resource content. However,

[1] http://www.nlm.nih.gov/mesh/

having the respective hierarchical resource available, it is also possible to automatically make an abstraction step from the possibly specific keywords to more general *topics* via ascension over the hierarchical paths. The main advantage of such topics over keywords is their conciseness, allowing to immediately detecting a sharp incongruence. Moreover, the aggregating effect may diminish the risk of isolated but strikingly irrelevant keywords unduly overturning the final opinion towards a negative result

Based upon these considerations, MeSH topic together with MeSH keywords represent one of the ten content attributes (criteria) currently used for 'intelligent' support of medical website certification in the MedIEQ project[2]. The selection of criteria was based on needs of two quality assessing agencies (Aqumed[3] and WMA[4]). The criteria are automatically extracted or inferred from web sites, 'subsites' (logically coherent portions of website) or documents, and then submitted to the agency's quality specialists in the form of pre-annotated and pre-classified data. The *topic classification* task comes along with MeSH term identification, and its main goal is to derive general topics from the set of identified MeSH keywords. From the Aqumed's point of view the initial motivation for determining the topic stems from their need to verify whether the web resource appropriately describes the goal of its own publication. The criterion can be verbalised by means of two questions: 'Is there an introduction that describes which topics are treated?' and 'Is there a detailed table of content?'. The other agency involved (WMA) only uses the website's topic (and target) for general resource description and categorisation.

Among other, the MedIEQ project's ambition is to deliver multilingual tools that would be able to process web resources within the multilingual European environment. In this respect, seven European languages were chosen for the project: Spanish, Catalan, German, English, Greek, Czech and Finnish.

The paper is structured as follows. In section 2 we present the general idea of concept matching and topic detection and the principles of our tools, in section 3 there is a concrete example of the classification task. Finally, we review related work in section 4 and future steps in section 5.

2. Problem Description and Methods Used

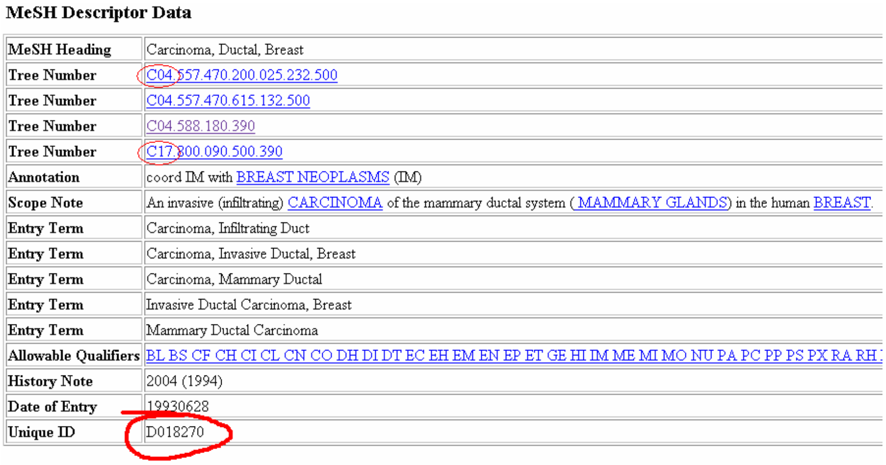
2.1. Structure of MeSH Concepts and 'Topics'

The MeSH is a polyhierarchy of medical concepts (i.e. single concept can exist in more than one place in different contexts). Within the hierarchy each meaning, in MeSH terminology called concept, is represented by its *Unique ID*, its *MeSH heading* (default name) and its *Tree Numbers* (contexts). The tree numbers are in fact all distinct drill paths leading from the root of the hierarchy to the concept. In addition to the Unique ID and contexts, each concept can be looked up by its *entry terms*, i.e. the list of concept's name and synonymous labels (see Figure 1).

[2] http://www.medieq.org

[3] http://www.azq.de

[4] http://wma.comb.es



MeSH Descriptor Data

MeSH Heading	Carcinoma, Ductal, Breast
Tree Number	C04.557.470.200.025.232.500
Tree Number	C04.557.470.615.132.500
Tree Number	C04.588.180.390
Tree Number	C17.800.090.500.390
Annotation	coord IM with BREAST NEOPLASMS (IM)
Scope Note	An invasive (infiltrating) CARCINOMA of the mammary ductal system (MAMMARY GLANDS) in the human BREAST.
Entry Term	Carcinoma, Infiltrating Duct
Entry Term	Carcinoma, Invasive Ductal, Breast
Entry Term	Carcinoma, Mammary Ductal
Entry Term	Invasive Ductal Carcinoma, Breast
Entry Term	Mammary Ductal Carcinoma
Allowable Qualifiers	BL BS CF CH CI CL CN CO DH DI DT EC EH EM EN EP ET GE HI IM ME MI MO NU PA PC PP PS PX RA RH
History Note	2004 (1994)
Date of Entry	19930628
Unique ID	D018270

Figure 1 – MeSH Descriptor Data view in MeSH browser [1], with demarcation for the concept's Unique ID and for the 'topic' level on the first position of tree numbers. The tree numbers themselves represent distinct contexts for the concept.

The MeSH hierarchy has been translated to several languages. In the MedIEQ project we decided to use the English, Spanish, German, Czech and Finnish variants of the MeSH vocabulary, as integrated in the UMLS metathesaurus database structure. In UMLS each of MeSH variant Unique IDs receives a UMLS unique atom ID (AUI) and the same concepts from different languages are aggregated under a common CUI (concept unique identifier). MeSH contexts are replaced by UMLS relationships and entry terms are transformed into UMLS word/string search structures. In addition, we transformed MeSH to SKOS[5] format for keyword extraction by methods defined in [6].

The term *topic*, by convention established within the MedIEQ project, refers to concepts on the 2nd most general level of the MeSH hierarchy (see Figure 1). On this level of hierarchy there are about 130 concepts. Each of them occurs in just one and only context. Each topic is then defined, same as other concepts, by its Unique ID (or its UMLS CUI respectively), the MeSH heading and the tree number. For example the topic 'Neoplasms[C04]' (Unique ID = D009369, CUI = C0027651) is defined as child of the topmost category Diseases[C].

2.2. Human Language Issues

The quality of MeSH concept/topic identification is in reality negatively influenced by issues related to human language ambiguity or by ambiguity resulting from our technical incapability to transfer human language to its computer representation properly.

2.2.1. Synonymy

The first feature of human language we need to cope with is *synonymy* together with *colloquiality*. The MeSH thesaurus was originally designed to allow browsing

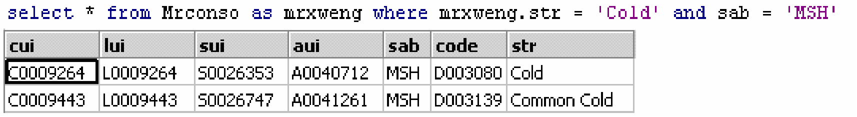
[5] http://www.w3.org/2004/02/skos/

correct medical terminology; it is in fact a nomenclature of valid terms in the medical domain. The canonical terms were later on supplemented by their most frequent synonyms as concept entry terms, but this process is far from being complete and universal. The English MeSH is the most elaborate version in this regard; nevertheless, there do not seem to be colloquial terms included. Therefore, when trying to cope with the reverse problem (i.e. matching widely used terms from the text to this hierarchy) the matching often fails simply because some synonyms and colloquial terms do not exist among entry terms of the concept.

Another measure negatively influenced by synonymy (this time from the perspective of whole document and its topic classification) is the frequency of matched term candidates (TC) in the document because we often fail to recognize that multiple terms belong to the same concept (see the example with *cold* and *common cold* in next subsection). A very common situation is that the author explains the meaning of the term in the beginning of the text and then just refers to it using the shorter form. Even worse, the shortened form is usually a colloquial term and the matching thus fails completely.

2.2.2. False Homonymy

In general the missing synonyms problem may be partially solved on the one hand by techniques allowing manual adjustment of the hierarchy (i.e. addition of required entry terms) which supports default *exact matching* algorithm or on the other hand by applying less strict matching algorithms. In contrary to matching exact terms only, *any match is good* approach brings new problems in form of wrong matches generating false homonymy: one word matches more than one concept even though they do not have the same entry term labels. As example of this effect see the following text: '*The COMMON COLD generally involves a runny nose, nasal congestion and sneezing. Over 200 viruses can cause a COLD*'. In MeSH there are two concepts containing the word 'cold' present. The first, *Common cold*, refers to the disease, and the other, *Cold,* refers to the temperature in the context of weather or physics. Even though we have two references to the disease in our sentence, neither of two approaches (*exact* or *any match*) leads to correct matching, since the concept *common cold* does not have the word *cold* as entry term. The *exact match* approach results only in matching *common cold* once from the first occurrence (COMMON COLD) and *cold* once from the second (COLD). On the other hand the *any match* approach results in matching both *common cold* and *cold* concepts for both occurrences. Unfortunately for the *any* match approach we also get plenty of other concepts such as 'cold fusion' or 'cold burn' as it matches all the concepts containing string 'cold' within its entry term. The most promising approach seems to be *exact match with synonym terms populated to the vocabulary*. When adding a synonym 'cold' to concept *common cold,* in *exact match we get* the result matching *common cold* twice and *cold* once. This way we get two correct and only one more wrong match in form of 'cold' as the temperature (see Figure 2).



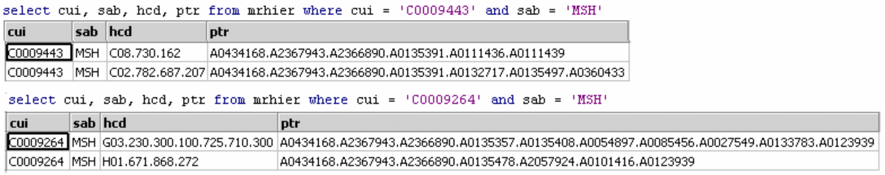
```
select * from Mrconso as mrxweng where mrxweng.str = 'Cold' and sab = 'MSH'
```

cui	lui	sui	aui	sab	code	str
C0009264	L0009264	S0026353	A0040712	MSH	D003080	Cold
C0009443	L0009443	S0026747	A0041261	MSH	D003139	Common Cold

Figure 2 – SQL query run against UMLS database showing two different concepts in Mrconso table that may become 'false homonyms' during MeSH term matching.

2.2.3. Thesaurus Contexts

The MeSH thesaurus is built so that many of its concepts exist in more than one occurrence (context). For example the concept *Common Cold* exists in two contexts (Virus Diseases [C02] and Respiratory Tract Diseases [C08]) and the concept *Cold* in another two different contexts (G03 and H01) (see Figure 3). Once such concepts are matched we need to take into account all of its contexts for the topic classification task, since the right context (same as right concepts) can only be resolved by understanding the semantics of the text.



```
select cui, sab, hcd, ptr from mrhier where cui = 'C0009443' and sab = 'MSH'
```

cui	sab	hcd	ptr
C0009443	MSH	C08.730.162	A0434168.A2367943.A2366890.A0135391.A0111436.A0111439
C0009443	MSH	C02.782.687.207	A0434168.A2367943.A2366890.A0135391.A0132717.A0135497.A0360433

```
select cui, sab, hcd, ptr from mrhier where cui = 'C0009264' and sab = 'MSH'
```

cui	sab	hcd	ptr
C0009264	MSH	G03.230.300.100.725.710.300	A0434168.A2367943.A2366890.A0135357.A0135408.A0054897.A0085456.A0027549.A0133783.A0123939
C0009264	MSH	H01.671.868.272	A0434168.A2367943.A2366890.A0135478.A2057924.A0101416.A0123939

Figure 3 – SQL query showing two different contexts for two candidate concepts (CC)

2.3. Automated Keywords and Topic Detection in MeSH

The overall process of detecting topic categories from HTML documents consists in three major steps.

The first task to do is content *pre-processing* that extracts the textual input from HTML for MeSH term matching. The pre-processed text contains *term candidates* (TC), which generally amounts to all consequent word combinations (1, 2 or 3-grams).

The subsequent *term matching* process takes term candidates from the input and matches them against MeSH entry terms or UMLS search structures [2] concepts yielding one or more *concept candidate* (CC) per successful TC. Within MedIEQ we consider employing two different matching algorithms at the moment. The first one, developed by the UNED[6], is based on extracting terms candidates from the text and matching them by pairs to the UMLS database [3] loaded with MeSH data using the UMLS MetamorphoSys tool [4] in order to obtain list of most probable *candidate concepts*. This algorithm only works with term candidates of maximum length of two words (2-grams). The second approach, developed by the TKK[7], uses the MeSH in SKOS format. The matching itself is done by a component called Poka [5]. In both cases, each term candidate can match several most probable concept candidates (CC). The output of the term matching step is text with highlighted term candidates for which the matching succeeded, together with the list of mapped concept candidates and their frequencies of occurrence within the document (see Figure 3).

The last step of the process is the *topic classification* itself. As discussed above by the term *topic* we understand concepts from the 2nd level of MeSH hierarchy. These concepts appear on the first position of the concept tree numbers. We can thus simply calculate the overall *score* of topic *T* for document *d*, denoted as *Score(T,d)*, from all occurrences of concepts from the set of *descendants* of T, *desc(T)*. For each concept *C*

[6] Universidad Nacional de Educación a Distancia, http://nlp.uned.es
[7] Helsinky University of Technology, http://www.tkk.fi

from *desc(T)* we have to consider all contexts (i.e. paths) with respect to *T*, *(paths(C,T))*, as well as the number of occurrences of *C* in document *d*, *freq(C,d)*. In the simple aggregation method we use, *Score(T,d)* is then calculated as the sum of products of *paths(C,T)* and *freq(C,d)* over all $C \in desc(T)$.

$$\text{Score}(T,d) = \sum_{C \in \text{desc}(T)} (\text{paths}(C,T) * \text{freq}(C,d))$$

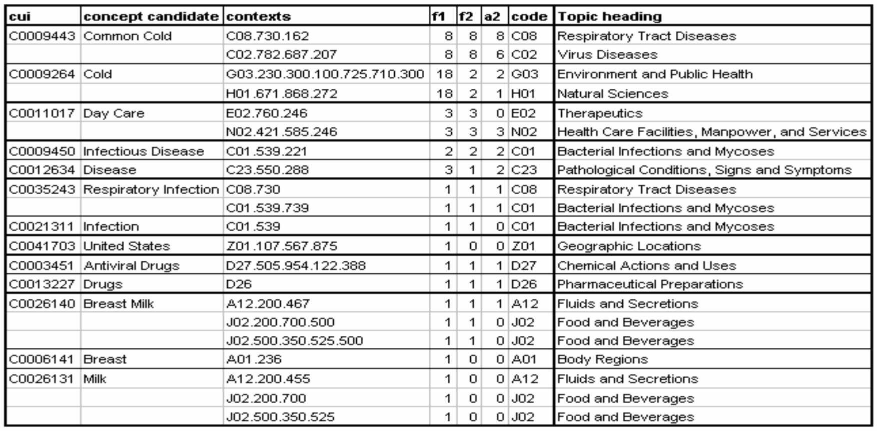
The output of the classification is a list of 'most probable' topic concepts ranked by their score. We use the ranking in order to get the top (at the moment top 10 with ties) most probable topics. The other reason for having the output ranked is that the comparison of efficiency between different approaches may be easily done based on it (see Figure 3).

3. Initial Experiments

The crucial for the quality of topic classification is the matching phase of the process together with the way concept candidates and their contexts are treated. In the current phase of the project we performed matching and topic identification experiments on English and Spanish texts. In addition the MeSH term matching was also tested on Finnish and Czech texts.

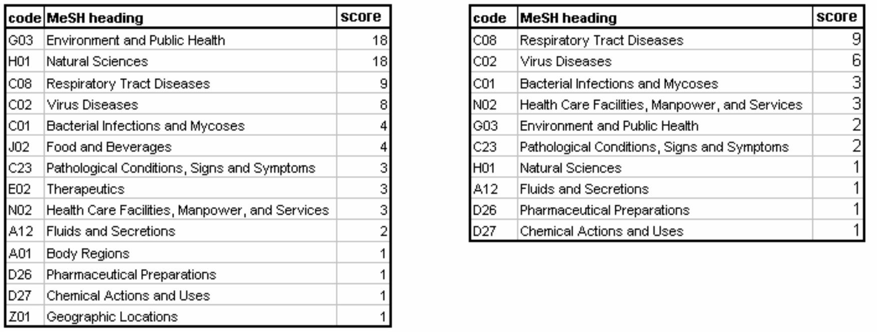
Two extreme approaches in terms of dependence between additional user effort required and the correctness of the topic classification have been analyzed. The first of them, a *simple automatic* approach, uses any of three matching techniques discussed in section 2.2.2 without any further user decision making (all of the concept candidates and all of their contexts are utilized). For any of matching techniques it produces relatively good topics on the top of the result list accompanied by quite a many wrong ones distributed all over the ranked list at the same time. The highest number of wrong topics is generated by *any match* strategy. The opposite extreme approach tested is based on semiautomatic matching-classification. The automatic output in form of document annotated by matched concept candidates and their contexts is offered to the user who performs the manual evaluation and selection of the right concepts and the right contexts in consequence. Such manual adjustment demonstrably leads to substantial increase in the quality of topic classification (see Figures 3 and 4).

From the user point of view there seems to be quite a big difference in labour intensity between both of *exact match* approaches and the *any match* approach as the latter yields lots of wrong matches that need to be thus manually examined.



cui	concept candidate	contexts	f1	f2	a2	code	Topic heading
C0009443	Common Cold	C08.730.162	8	8	8	C08	Respiratory Tract Diseases
		C02.782.687.207	8	8	6	C02	Virus Diseases
C0009264	Cold	G03.230.300.100.725.710.300	18	2	2	G03	Environment and Public Health
		H01.671.868.272	18	2	1	H01	Natural Sciences
C0011017	Day Care	E02.760.246	3	3	0	E02	Therapeutics
		N02.421.585.246	3	3	3	N02	Health Care Facilities, Manpower, and Services
C0009450	Infectious Disease	C01.539.221	2	2	2	C01	Bacterial Infections and Mycoses
C0012634	Disease	C23.550.288	3	1	2	C23	Pathological Conditions, Signs and Symptoms
C0035243	Respiratory Infection	C08.730	1	1	1	C08	Respiratory Tract Diseases
		C01.539.739	1	1	1	C01	Bacterial Infections and Mycoses
C0021311	Infection	C01.539	1	1	0	C01	Bacterial Infections and Mycoses
C0041703	United States	Z01.107.567.875	1	0	0	Z01	Geographic Locations
C0003451	Antiviral Drugs	D27.505.954.122.388	1	1	1	D27	Chemical Actions and Uses
C0013227	Drugs	D26	1	1	1	D26	Pharmaceutical Preparations
C0026140	Breast Milk	A12.200.467	1	1	1	A12	Fluids and Secretions
		J02.200.700.500	1	1	0	J02	Food and Beverages
		J02.500.350.525.500	1	1	0	J02	Food and Beverages
C0006141	Breast	A01.236	1	0	0	A01	Body Regions
C0026131	Milk	A12.200.455	1	0	0	A12	Fluids and Secretions
		J02.200.700	1	0	0	J02	Food and Beverages
		J02.500.350.525	1	0	0	J02	Food and Beverages

Figure 3 – Comparison of topic candidate frequencies freq(C,T) for two extreme approaches. Column f1 indicates frequency for all the matched candidate concepts used in the 'simple automatic approach' assigned to all its available concepts. Column f2 shows frequency values for accepted concepts only (after manual selection of the right concepts) and column a2 refers to frequencies for further filtered/accepted contexts only.



code	MeSH heading	score
G03	Environment and Public Health	18
H01	Natural Sciences	18
C08	Respiratory Tract Diseases	9
C02	Virus Diseases	8
C01	Bacterial Infections and Mycoses	4
J02	Food and Beverages	4
C23	Pathological Conditions, Signs and Symptoms	3
E02	Therapeutics	3
N02	Health Care Facilities, Manpower, and Services	3
A12	Fluids and Secretions	2
A01	Body Regions	1
D26	Pharmaceutical Preparations	1
D27	Chemical Actions and Uses	1
Z01	Geographic Locations	1

code	MeSH heading	score
C08	Respiratory Tract Diseases	9
C02	Virus Diseases	6
C01	Bacterial Infections and Mycoses	3
N02	Health Care Facilities, Manpower, and Services	3
G03	Environment and Public Health	2
C23	Pathological Conditions, Signs and Symptoms	2
H01	Natural Sciences	1
A12	Fluids and Secretions	1
D26	Pharmaceutical Preparations	1
D27	Chemical Actions and Uses	1

Figure 4 – Output ranked (by score(T,d)) topic lists obtained by two extreme approaches (left: 'simple automatic' approach; right: semiautomatic approach with manual selection of the right concepts/contexts)

4. Related Work

There are quite a number of tools and projects covering the MeSH term extraction task either as standalone functionality or just as a component of more complex systems. The NLM – the provider of the UMLS database (including MeSH) itself offers a tool called MMTx (MetaMap Transfer) [7] intended for this purpose. It is based on UMLS and profits from its elaborate search structures. Unfortunately these structures are easily usable for English only. The leading organization in the field of assessing quality of web resources HON developed for its purposes the tool MARVIN (Multi-Agent Retrieval Vagabond on Information Network) [8], which extracts MeSH terms while indexing pages. Based on the extracted sets it constructs an inverse index that can be searched by HON's search tools later on. Another project employing MeSH term matching is CISMeF [9, 10]. It specializes on francophone web resources, and thus uses the French version of the MeSH hierarchy. A common feature of all these tools is that they perform very well on one language but they are very uneasy to extend their functionality to other languages. Since the MedIEQ project intends to cover a wide range of languages (many of which were not treated before in context of MeSH concepts) with a unified approach, we had to develop our own functionality coping with this task had to be developed.

However the term *MeSH topic* was established for internal project requirements, the need for finding some more general document topics or categories (e.g. target audience) is mentioned in few other projects. Relatively relevant to our topic identification task seems to be the work done on a MEDLINE categorization algorithm in CISMeF [11]. Their algorithm is based on semantic links between MeSH terms and meta-terms on the one hand and between MeSH subheadings and meta-terms on the other hand. This algorithm is used to categorize scientific articles in MEDLINE indexed by MeSH terms. Primarily it was conceived to improve the recall of search queries.

5. Conclusions and Future Work

We discussed the problem of keyword and topic detection in medical websites, and explained the approach developed for this problem in the context of the MedIEQ project. Although relatively inconspicuous, the keyword and topic detection task is quite important in the overall workflow of website certification.

The result of topic classification based on fully automatic classification approach in comparison to semiautomatic approach shows that we are possibly able to achieve quite meaningful results on the top positions of the topic list even without manual user intervention; however the output tends to become degraded by topics of poor relevance mainly at the bottom of the ranked list.

Great improvement in quality of topic classification is achieved in semi-automatic process as professional quality assessment users may easily pick the right concepts and their right contexts. The ultimate goal of the future project effort lies thus in shifting as much of work load as possible towards the automatic part of the process. In matching phase there could be accomplished some improvement using ad-hoc thesaurus adjustments (adding new entry terms), however it might bring some future compatibility problems during ontology updates. Some success in automatic disclosing of the right concept candidates and the right contexts might be possibly achieved either by analysis of nearby word occurrences or by analysing the context co-occurrence among different matched concept candidates. Co-occurring contexts would then get some extra score points in contrast to those where the context seems to be unique within the document. Another potentially effective way to get rid of irrelevant topics would be to focus on categories closely related to the medical domain (e.g. including *diseases* [C] and *chemicals and drugs* [D] and leaving off *geographicals* [Z]), in other words, to choose relevant subsets only. The question remains whether to apply restrictions at the level of the indexing/spidering process (which precedes the actual document analysis), of the matching process, of the output topic list or maybe at all of these levels or combination of them.

At this stage of the project we carried out thorough experiments with English and Spanish, and initial experiments with some other languages. We expect new issues to appear mainly in matching phase (as topic classification itself seems to be language independent) as approaching in-depth other languages such as Finnish or Czech, where the MeSH variant is unsophisticated and mainly restricted to scientific terminology, unlikely to perform well in the matching task. Approaches for semiautomatic extension the national versions of the thesaurus, such as outlined in [12], would be worth investigating. Furthermore languages such as Catalan which not to have any existing MeSH translation at the moment need to be tackled similarly.

The approach introduced in this paper can be adapted on different domains if exists a closed vocabulary with a structure with following terms. 1) Upper level concepts comprise a relevant and categorizing set of concepts on the subject matter. 2) The lower level concepts occurrences in text indicate the existence of the topic.

6. References

[1] http://www.nlm.nih.gov/mesh and http://www.nlm.nih.gov/mesh/MBrowser.html

[2] http://www.nlm.nih.gov/research/umls/meta2.html#s2_6

[3] http://www.nlm.nih.gov/research/umls

[4] http://www.nlm.nih.gov/research/umls/meta6.html

[5] Valkeapää O, Alm O, Hyvönen E: **An Adaptable Framework for Ontology-based Content Creation on the Semantic Web.** *Journal of Universal Computer Science*, vol. 13, no. 12 (2007), 1835-1853

[6] van Assem M, Malaise V., Miles A, Schreiber, G..:**A method to Convert Thesauri to SKOS.** *3rd European Semantic Web Conference (ESWC2006).*

[7] http://www.nlm.nih.gov/research/umls/mmtx.html

[8] 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

[9] Baud RH, Ruch P, Gaudinat A, Fabry P, Lovis C, Geissbuhler A: **Coping with the variability of medical terms.** *Medinfo.* 2004;11(Pt 1):322-6

[10] http://www.cismef.org and http://www.chu-rouen.fr/cismef

[11] Darmoni SJ, Névéol A, Renard JM, Gehanno JF, Soualmia LF, Dahamna B, Thirion B: **A MEDLINE categorization algorithm.** *BMC Med Inform Decis Mak.* 2006 Feb 7;6:7.

[12] Kolesa P , Přečková P : **Tools of Czech Biomedical Ontologies Creation. In: Ubiquity: Technologies for Better Health in Aging Societies. (Ed.: Hasman A., Haux R., van der Lei J., De Clercq E., France F.H.R.) - Amsterdam**, *MIE 2006. International Conference of the European Federation for Medical Informatics* /20./, Maastricht, IOS Press, 2006, pp. 775–80.