

Accessing Concepts in Medical Terminology Resources

Manuel Fidalgo^a, Enrique Amigó^a and Felisa Verdejo^a

^a*Dept. Lenguajes y Sistemas Informáticos de la UNED, Spain*

Abstract. This document is focused on the interactive access to concepts in terminological resources (UMLS). The goal is to analyse which query processing techniques are necessary to offer a ranking containing the required concept in first positions. In order to this, a simple approach is compared against other current approaches based on linguistic processing.

Keywords. Medical domain, terminological resources, UMLS, customisation, browsing.

Introduction

A wide set of terminological resources has been developed, containing information about biomedical and health related concepts, their various names, and the relationships among them. Some examples are the Medical Subject Headings (Mesh) [3], SNOMED [2] or ICD-9-CM [4]. In the context of the National Library of Medicine's Unified Medical Language System (UMLS) project, a metathesaurus covering more than one hundred resources has been developed mapping concepts between resources [1].

UMLS includes currently a wide set of terminological resources focused on multiple specific domains. Therefore, for a particular task it is convenient and efficient to select a relevant subset to the needed coverage. In order to customise UMLS for specific applications, each UMLS release provides the installation and customisation program *MetamorphoSys*. This program includes the following functionalities: selecting text formats, excluding specific languages, eliminating relationships, eliminating all concepts assigned to a particular semantic type. However, the management of these functionalities requires a computer expert profile. The motivation of this work is to provide a user-friendly interface allowing medicine experts enhanced functionality to manage UMLS in order to generate a customised collection.

For this purpose three basic functionalities are necessary: (1) accessing specific concepts, (2) browsing across relationships and (3) selecting concept sub-hierarchies in order to generate customised resources. This work is focused on the access to specific concepts by means of user queries. The most popular tool for mapping medical concepts from texts is the *MetaMap Transfer tool (MMTx)*¹. However, it requires linguistic processing which is language dependent while terms from multiple languages are available in

¹<http://mmtx.nlm.nih.gov/>

- terms searched: common cold
- Total concepts found: 835
- Took: 5 seconds

Search

Enter the terms to search:
common cold

Results per page: 10 Language ENG Vocabulary ALL Search

Results

AUI	CUI	Concept	Vocabulary	Language	Show concept
A4739976	C0009443	Common cold	SNOMEDCT	ENG	Show concept
A0041261	C0009443	Common Cold	MSH	ENG	Show concept
A7757440	C0205939	Common Cold Virus	MSH	ENG	Show concept
A2914644	C0199822	Common cold vaccination	SNOMEDCT	ENG	Show concept
A3267154	C0035473	Common cold virus group	SNOMEDCT	ENG	Show concept
A3085519	C0851143	[V]Common cold vaccination	SNOMEDCT	ENG	Show concept
A2992311	C0161636	Anti-common cold drug poisoning	SNOMEDCT	ENG	Show concept
A3649902	C1299230	Proprietary compound common cold remedy	SNOMEDCT	ENG	Show concept
A3092934	C0414040	Adverse reaction to anti-common cold drugs	SNOMEDCT	ENG	Show concept
A3239375	C0568722	Accidental anti-common cold drug poisoning	SNOMEDCT	ENG	Show concept

Figure 1. MRM concept retrieval interface

UMLS terminological resources. In order to avoid the need for language dependent linguistic tools, in this paper we analyse which linguistic processes are necessary to obtain an acceptable concept ranking.

This work is placed into the context of the Multilingual Resources Manager (MRM), a system developed in the context of the MedIEQ project². The rest of the paper is organised as follows. Section 1 describes briefly the MRM system. Section 2 compares the task tackled in this article with other similar tasks and enumerates other current approaches to our problem. Section 3 defines the evaluation framework for comparing MRM concept access method against other current approaches. Results are discussed in Section 4, and the conclusion section focuses on contributions and further work.

1. Multilingual Resources Manager system (MRM)

The main MRM goal is to allow medical experts to pick concepts from UMLS terminological resources, hiding data model and implementation aspects. In the context of the AQUA system (MedIEQ project³) these picked concepts are employed for crawling medical web pages and other tasks.

Figure 1 shows the MRM user interface for accessing specific concepts. The search box in the figure shows the user query “common cold”, the selected language (English) and the resources to be considered (ALL). The query contains the term “common cold” and it is processed in the following way: (1) Phrases in the selected terminological re-

²<http://www.medieq.org/>

³<http://www.medieq.org/>

Concept information

Name	Cui	Vocabulary	Language	Semantic type	Definition
Asthma	C0004096	MSH	ENG	Disease or Syndrome	A form of bronchial disorder associated with airway obstruction, marked by recurrent attacks of paroxysmal dyspnea, with wheezing due to spasmodic contraction of the bronchi.

Synonyms: [Asthmas](#)

Related concepts: [Bronchial Hyperreactivity \(MSH\)](#) [Anti-Asthmatic Agents \(MSH\)](#)

Select resource to store the concepts: [Create resources](#)

Show hierarchy in: [English](#) [Spanish](#)

Select context:

Hierarchy

MeSH Descriptors
[Index Medicus Descriptor](#)
[Diseases \(MeSH Category\)](#)
[Respiratory Tract Diseases](#)
[Bronchial Diseases](#)
Asthma
[Asthma, Exercise-Induced](#)
[Status Asthmaticus](#)
[Bronchial Fistula](#)
[Bronchial Hyperreactivity](#)
[Bronchial Neoplasms](#)
[Bronchial Spasm](#)
[Bronchiectasis](#)
[Bronchitis](#)
[Bronchogenic Cyst](#)
[Bronchopneumonia](#)
[Tracheobronchomegaly](#)

Figure 2. MRM concept browsing interface

sources containing at least one query word are collected. Query words can be included into concept words. In order to avoid an excessive amount of matches, words shorter than three characters are not considered. (2) Retrieved phrases are ranked according to the number of query words in each phrase and the minimum phrase length. We have realised that shorter terms tend to avoid over-specified concepts. (3) Phrases related to the same concept are removed from the list. (4) The final list of concepts is displayed according to the ranking criteria previously described.

The interface displays for each concept the related phrase (third column) the corresponding resource (vocabulary, fourth column), the language and a link to the browsing interface to show further contextual information. Figure 2 shows the MRM interface for visualising a concept description and for browsing across relationships. The buttons “add concept” and “add sub-hierarchy” allow the user to include new concepts in the customised resource. Showing a ranking list of concepts for a user query has been considered in different approaches. However, there is not any evaluation about the ranking criteria in this specific task. The rest of the paper focuses on this problem.

2. Accessing concepts

We find in the literature two kinds of tasks related to our problem: Extracting medical concepts from full texts and mapping automatically concepts across ontologies. Next we will elaborate a comparison.

A substantial amount of research on concept recognition within clinical texts has been done [8,9,10]. The most popular tool for mapping medical concepts from texts is

the MetaMap Transfer tool (MMTx)⁴. This tool uses a pipeline of linguistic processes including generating phrase variants and synonym expansion. In order to rank candidates a function combining the four following criteria is applied: (1) the proportion of query words in the retrieved concept term, (2) the linguistic variation between the query and the candidate term, (3) the centrality of the query in the retrieved term and (4) the cohesiveness between query words. In our case, medical concepts have to be identified from a query expression, rather than a full text. This implies that some of the linguistic preprocessing performed with MMTx would be not necessary for our purposes. So the question to be further investigated is the kind of linguistic preprocessing needed for short queries.

Other works have focused on automatic mapping of concepts between terminological resources [11]. This problem is very related to our case. However, a difference is that we provide a ranking of candidates. Therefore we are not looking for a unique solution (yes or not mapping) but for a list of potential ones, appropriately ranked. This implies again that some linguistic processes could be avoided obtaining similar results in terms of user needs. The criteria for evaluating a mapping do not apply to evaluate rankings and therefore further investigation is needed on this topic.

In the context of our task, some current web services such as the MESH browser⁵ or the SNOMED CaTTs browser⁶ provides concepts rankings using non language dependent strategies. The MESH browser returns in alphabetical order all terms in MESH containing at least one query word. CaTTs browser applies similar criteria as MRM, but returning just terms containing all query words.

3. Evaluation framework

An important aspect to evaluate an information retrieval system is to have available a set of user queries and relevance judgements for the system outcomes. Developing such a testbed is considerably expensive. However, in our case we can extract from UMLS resources the necessary information to automatically generate a testbed.

Each resource in UMLS contains a wide set of medical concepts, each one associated with a set of phrasal terms. Concept identifiers are also mapped between resources in UMLS. For instance, the concept C0007680 is associated with the terms “Central Nervous System Agents”, “CNS agents” and “Central Nervous System Drugs” in MESH. In addition, the same concept identifier is associated with the terms “Central nervous system agent (substance)”, “CNS drug” and “Central nervous system agent (product)” in SNOMED. We consider that these human annotated terms represent potential user queries in the context of the MRM accessing task.

The evaluation procedure consisted on searching in MESH the corresponding concept using a term provided by SNOMED. On one hand we used the MESH browser, MRM and MMTx for searching concepts in MESH using SNOMED query terms considering 74 randomly selected samples. On the other hand, we used SNOMED CaTTs browser, MRM and MMTx for searching concepts in SNOMED using MESH terms considering 90 samples. The evaluation measure is the correct concept rank. Ranks higher than 15 were considered as “non found” concepts while ranks lower than 15 were considered as “found” concepts.

⁴<http://mmtx.nlm.nih.gov/>

⁵<http://www.nlm.nih.gov/mesh/MBrowser.html>

⁶<http://www.jdet.com/>

4. Results

First of all, our experiment has shown that MESH browser can not be considered a baseline for evaluation purposes. For instance, the MESH browser returns the correct concept among the first 15 positions just in 8 cases from 74, while MRM in 50 cases. This is due to the fact that the MESH browser displays the results in alphabetical order. MRM improves also the SNOMED CaTTs browser, since the latter requires finding all query words in the candidate term. In order to smooth this effect we have removed iteratively the last query word. In spite of this, 37 from 95 samples were not retrieved at all by the CaTTs browser while just 16 cases were not retrieved by MRM.

We have compared as well MRM against MMTx tool and we have identified the reasons for non found concepts on both approaches. Table 1 shows the number of cases in which each approach retrieves the corresponding concept in a rank lower and higher than 15. Results indicate that MMTx performs better than MRM in both cases, suggesting that linguistic processing help to improves the accuracy. Then, the next issue is to study what linguistic processes should be adopted to process short queries for obtaining acceptable rankings.

Table 2 shows the main reasons of failures in both approaches. First, although MRM considers all related terms of candidate concepts, it does not expand synonyms independently from individual query words. Consequently, MRM failed in 41 cases due to synonyms while MMTx just failed in 10 cases. However, the MRM approach considers partial matching of words, solving all abbreviations while MMTx failed in 23 cases for this category. MMTx can not consider partial matching of words because for full documents this would generate too much noise. The lower limit for partial matching was established at three characters. This implies non considering words shorter than 3 characters. This was the reason for 2 failures. Finally, non considering apostrophes in MRM produced 3 failures.

In conclusion, our results suggest that an appropriate approach for dealing with our task is to use MRM ranking criteria but including synonym expansion for individual words and apostrophe processing for the query. It is worth to notice that UMLS include synonyms. Therefore, synonym expansion can be done without additional resources.

5. Conclusions

As far as we know, although there exists several web services offering access to concepts in medical linguistic resources, the problem of accessing specific concepts through query processing has not been analysed in deep. In this paper we have presented the MRM system which offers this functionality, comparing it against other current approaches. Our evaluation framework based on the UMLS information allows to compare approaches without needing additional manual annotations. Our analysis had shed some lights about the key linguistic aspect to be taken into account in order to tackle the task. Synonym expansion seem to be a promising technique. The immediate future work consists on implementing this feature in the system and check the results using real users in the evaluation framework.

	MESH query term against SNOMED resource	SNOMED query term against MESH resource
MRM rank < 15	71	47
MMTx rank < 15	74	56
MRM rank > 15	27	24
MMTx rank > 15	28	11

Table 1. Retrieved concepts using MESH/SNOMED query terms over SNOMED/MESH resources. MRM and MMTx are compared.

Cause	MRM failures	MMTx failures
Synonyms	41	10
Abbreviations	0	23
Short words	2	0
Apostrophe	3	0
Non identified	5	6

Table 2. Linguistic causes of failures in concept retrieval approaches

Acknowledgements

This work has been supported by research grant MEDIEQ. We are indebted to three anonymous reviewers for their comments on earlier versions of this work

References

- [1] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. *Meth Inform Med.* 1993;32: 281-91. [PubMed].
- [2] International Health Terminology Standards Development Organisation, SNOMED Clinical Terms. January 2008 IHTSDO, Rued Langgaards Vej 7, 5.
- [3] Medical Subject Headings (Mesh);National Library of Medicine;2008;Bethesda (MD);<http://www.nlm.nih.gov/pubs/factsheets/mesh.html>
- [4] U.S. Department of Health and Human Services, Centers for Medicare & Medicaid Services;ICD-9-CM [computer file]: International Classification of Diseases, Ninth Revision, Clinical Modification;October 1, 2006;Baltimore, MD
- [5] Wagner MM, Cooper GF. Evaluation of a Meta-1 automatic indexing method for medical documents. *Comput Biomed Res.* 1992;25: 336-50. [PubMed]
- [6] Hersh WR. Evaluation of Meta-1 for a concept-based approach to the automated indexing and retrieval of bibliographic and full-text databases. *Med Decis Making.* 1991;11: suppl: S120-S124. [PubMed].
- [7] Powsner SM, Miller PL. Automated online transition from the medical record to the psychiatric literature. *Meth Inform Med.* 1992;31: 169-74. [PubMed].
- [8] Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. *J Am Med Informat Assoc.* 1994;1: 161-74.
- [9] Haug P, Koehler S, Lau LM, Wang P, Rocha R, Huff S. A natural language understanding system combining syntactic and semantic techniques. *Proc Symp Comput Appl Med Care.* 1994: 247-51.
- [10] Evans DA, Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proc Fall Symposium of the American Medical Informatics Association.* 1996: 388-392.
- [11] Cimino JJ, Johnson SB, Aguirre A, Roderer N, Clayton PD. The Medline button. *Proc Symp Comput Appl Med Care.* 1992: 81-5.