MEDINFO 2004
M. Fieschi et al. (Eds)
Amsterdam: IOS Press

WRAPIN: New Generation Health Search Engine Using UMLS Knowledge Sources for MeSH Term Extraction from Health Documentation

Arnaud Gaudinat[a], Michel Joubert[b], Sylvain Aymard[b], Laurent Falco[b], Célia Boyer[a], Marius Fieschi[b]

[a] *HON Foundation, Geneva, Switzerland*
[b] *LERTIM, Faculté de Médecine, Université de la Méditerranée, Marseille, France*

Abstract

To realize the potential of the Internet as a source of valuable healthcare information, for the general public, patients or practitioners, it is imperative to establish a validation system based on standards of quality. The WRAPIN project (Worldwide online Reliable Advice to Patients and Individuals) from the European Community has this ambitious goal. WRAPIN is a federating system for medical information with an editorial policy of intelligently sharing quality and professional information. The WRAPIN project has two main axes: the efficient and intelligent search of information and the assertion of the trustworthiness of content.

This article presents the scientific challenges involved in extracting the knowledge from text-based information in order to better manage the knowledge and the rest of the retrieval process. Our innovative approach is to efficiently extract MeSH terms from the analyzed documents exploiting UMLS knowledge sources[1]. A benefit has been measured when comparing extraction results. Even if the evaluation is made with a limited corpus, this research work proposes heuristics that can be validated to the whole biomedical domain, and possibly enhanced by the adjunction of other methods.

Keywords:

Trustworthiness; MeSH mapping; Information retrieval; UMLS; Co-occurrences.

Introduction

This work has been done in the framework of the European project WRAPIN (Worldwide online Reliable Advice to Patients and INdividuals)[2]. As expressed in this acronym, the aim of WRAPIN is to empower the citizen to judge the information he/she reads on the Web. This is possible by offering privileged access to trustworthy medical and health sources such as scientific articles from MEDLINE (by the National Library of Medicine[3]), OESO[4] and URO France[5], HONcode-accredited web sites as well as those described in HON's MedHunt database, found in Clinical Trials, FDA-provided news or HON news. Indeed, the search for specific information among the maze of documents available on the Web, especially in the medical domain, is a continual challenge for surfers since the Web's introduction. Many services now facilitate access to scattered resources, such as directories (beginning with Yahoo!), or specialized search engines like HealthFinder of the "U.S. Department of Health and Human Services", HealthLink, CliniWeb from "the Oregon Health & Science University", MEDLINEplus, from the "US National Library of Medicine", MedHunt and HONselect of "Health On the Net"[6]. Pioneering the medical and health Internet field since 1996, HON Foundation promotes the efficient use of the Internet for health matters by the citizen and by medical professionals. For this purpose HON developed the robot MARVIN (Multi-agent Retrieval Vagabond on Information Network) [1], partially funded by the Swiss National research fund, in order to retrieve and automatically index medical and health documents on the Web. In 1996, HON initiated the HONcode to qualify health resources on the Internet according to its initiated ethical standard [2]. Following these efforts, the WRAPIN project aims to enhance the results obtained both by: 1) improved indexing, retrieval and processing of health resources on the Internet, and 2) integration and dissemination of internationally renowned registries using the same indexing and retrieval processes.

To reach its objectives efficiently, WRAPIN had to enhance the capabilities of HON's retriever and indexer, MARVIN. This robot was conceived to screen the Web, retrieving only medical and health documents from the mass of documents available online. Once detected as pertaining to the medical and health field, the documents are analyzed and the content is fully indexed. The indexation is enriched, combining full text indexing with MeSH [3] indexing. The latter is composed of two steps. During the first it extracts the MeSH terms and matches any medical terms found to MeSH terms, and creates an inverted index. This index is searched by HON's own search tools, MedHunt and HONselect (freely available to the Internet community). Enhancements were judged to be required. This has been done with respect to results obtained in the framework of the ARIANE project [4]. The aim of this project was to provide users with seamless access to biomedical information from heterogeneous resources. De-

[1] Unified Medical Language System by the National Library of Medicine. http://www.nlm.nih.gov/pubs/factsheets/umlsmeta.html
[2] http://www.wrapin.org
[3] http://www.nlm.nih.gov
[4] Medical scientific articles about esophagus diseases
[5] French medical scientific articles in the urology domain.

[6] http://www.healthfinder.gov/, http://www.healthlinks.net/, http://www.ohsu.edu/cliniweb/,http://search.nlm.nih.gov/medlineplus/ http://www.hon.ch/MedHunt/, http://www.hon.ch/HONselect/

velopments of ARIANE were based on the UMLS (Unified Medical Language System) knowledge sources [5] provided by the National Library in Medicine. Specifically, ARIANE exploits the medical meaning declared in the UMLS Semantic Network [6].

The basic material that WRAPIN uses to index and retrieve documents is the MeSH terminology. So, the first challenge is to efficiently extract MeSH terms from the documents under analysis. The present work aims to show how the UMLS knowledge sources, especially the co-occurrences of terms in the MEDLINE literature, may contribute to a better indexation of medical documents making use of MeSH terms.

After a brief summary of the UMLS knowledge sources and methods used for extracting MeSH terms from documents, we will compare the results obtained by an extraction process before and after the integration of this kind of knowledge. Preliminary experiments have been made in the gastroenterology domain.

Materials and Methods

UMLS knowledge sources

The UMLS knowledge sources are partly made up of a Metathesaurus that integrates most of the nomenclatures used in the biomedical domain, and a Semantic Network. An important component of the Metathesaurus is the MeSH, which is available in fifteen languages in UMLS. The core concepts that have been isolated in the Metathesaurus are attached to generic types of concepts in the Semantic Network. These types are interconnected by directed binary semantic relationships. In this way, concepts may be potentially connected by means of semantic relationships that apply to types they are attached to.

Another UMLS source of knowledge is the table of co-occurring terms established mainly from MEDLINE. MEDLINE co-occurrence data are computed for concepts that were designated as principal or main points (with * attached to the main heading or any of the subheadings) in the same journal article. Each line describes a pair of (term/sub-headings) with a frequency. There are two rows in this table for each pair of co-occurring concepts, with one for each direction of their association, whose attributes and frequencies may vary. A previous work has shown how to exploit this table to translate a semantic relationship between two concepts into sub-headings attached to their designations in MeSH [7]. In our case we exploit this table in the inverse way. And conversely, a module computing the semantic relations, named "the ARIANE module" in the following, computes the translation of a pair of (term/sub-headings) into concepts connected with semantic relationships. More precisely, the module receives, as an input, two MeSH terms and it retrieves their associations in the table of co-occurring terms. Retrieving such associations, it computes the frequency of associations with respect to semantic relationships according to a table of correspondence between relations and subheadings. The module produces as an output a list of scored semantic relationships that apply to the two concepts that the input terms represented. Let us keep in mind that semantic relationships are directed relations that apply from one concept to another. Then, frequencies that are expressed on one direction are not equal to those directed conversely. That is why the module computes a score that cumulates the direct and inverse frequencies with respect to concepts and relations. The score calculated for a relationship is the sum of the two frequencies in direct and reverse directions. For instance, applying the process to "bile duct diseases" and "cholangiography" produces the following result: "cholangiography" is attached to the type "Diagnostic Procedure"; "bile duct diseases" is attached to the type "Disease or Syndrome". The scores obtained for each association (semantic relationship) are:

- diagnoses: 21
- measures: 11
- affects: 2

An extract of the knowledge from which these scores are computed is, for the period 1997-2001, from "bile duct diseases" to "cholangiography" is:

- Diagnosis: 6
- Radiography: 6
- Etiology: 2
- Therapy: 2
- Physiopathology: 1
- Surgery: 1

Knowing that several subheadings may be represented in a single line of the table, explains why the totals may be different.

What is noteworthy is that co-occurrences serve to give a weight to the relationships that apply on concepts. For instance, the above subheadings "Diagnosis" and "Radiography" intervene strongly in the computation of the score affected to the relationship "diagnoses", the score of which is greater than the score affected to the relationship "affects". This mechanism is a way to instantiate parts of the Semantic Network of the UMLS and allocates weights to the involved semantic relationships.

Extraction methods

An great amount of research on concept recognition in medical text has already been done by researchers [8-12]. The best known is probably the indexing initiative from the U.S. National Library of Medicine that uses three kinds of methods. This work has resulted in the well-known MetaMap [13, 14] system. The usual approach is mainly lexical, but several good results have been obtained with stochastic methods. Recently, we also investigated a non-supervised approach based on a space vector model with results as good as those obtained with MetaMap for a MeSH term extraction task. This research work has shown this method to be entirely adequate when working with a relatively large number of classes (more than 30,000 in case of MeSH). In fact, all these methods have benefits and drawbacks, but surprisingly they rarely use medical knowledge such as that available in the UMLS knowledge sources.

Our goal here is to use the UMLS semantic information described previously to enhance the MeSH term extraction task and, more particularly, to increase the precision of the system. This is a major challenge of all indexing and retrieval systems. In comparison with other techniques, which use linear combinations of the score to increase precision for a given acceptable recall, we use the semantic relationships between concepts via the co-occurrence list. In fact we have made the assumption that more relations between them characterize important concepts. We can call this hypothesis "the related term heuristics". Taking into account this hypothesis happens once the MeSH extraction is achieved, in reranking candidates (i.e. MeSH terms which best characterize the analyzed document). For that we proceed in two steps, 1) for each pair of terms in the terms list, we calculate the cumulate weight of their relations, and 2) we express each term according to its related pairs to obtain a unique score by term.

Let m_1, m_2, .. m_n the list of terms identified by the MeSH extractor system, where n is the number of MeSH terms.

Thus, for each combination of terms $(m_{i,} m_j)$ the ARIANE module (described in the previous section) can yield several relations with their associated scores. The cumulated scores $sc(m_{i,} m_j)$ of these relations give a weight to the relations which exist between the term m_i and m_j.

With a score sc for each pair of terms, we can then determine a score for each term, in relation to all pairs containing that term, except of course for the pair (m_i, m_i) which would be meaningless.

So the average score st for a MeSH term m_i is given by (1):

$$st(m_i) = \frac{\sum_{i=1, i \neq j}^{n} sc(m_i, m_j)}{(2n-2)}$$

The figure 1 illustrates an example of three MeSH terms m_1, m_2 and m_3 candidates. For each pair of candidates, there exist two possible relations represented by two arrows. For each relation a score SC has been calculated from UMLS by the ARIANE module. With these six scores it is easy to express the weight of one MeSH term relative to others. The equation (2) gives the weight st of the word m_2:

$$stm_2 = sc(m_2, m_1) + sc(m_1, m_2) + sc(m_2, m_3) + sc(m_3, m_2) \qquad (2)$$

Table 1 is an example of the list of MeSH terms obtained from a MEDLINE abstract and re-ranked automatically by our system. This list is compared to the original annotated list[1] (by the NLM experts) given in its citation. The MeSH terms shared by our system and the annotated[2] list are in gray.

The first column give the rank of the MeSH terms in decreasing order of interest, the second gives the candidates list obtained by our basic extraction system (our base system is described in the next section) and finally the third column gives the same candidates list re-ranked with respect to "the most related term heuristic" using UMLS knowledge sources exploited by the ARIANE module. As one can see, the basic extraction system gives five false MeSH terms in first followed by six good candidates, whereas our improvement with UMLS allows selecting in the first place five good candidates. Even if the term "Therapeutics" appears in tenth position, in comparing just the ranking, the precision[3] improvement is 32.5% with the usage of UMLS. It is important to underline that the precision, and then the ranking, are crucial when we want, for example, to query MEDLINE. In this case, the usage of twenty searching drives MEDLINE to silence and so the selection of the five best searching keys among the twenty becomes essential.

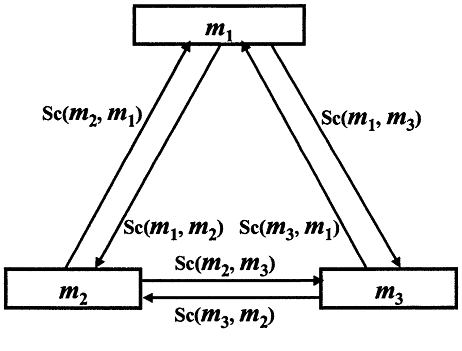

Figure 1 - Relations between three MeSH terms

Experimental Methods

The evaluation has been conducted on a corpus only extracted from MEDLINE. The MEDLINE database allows us to easily obtain a list of all the MeSH terms occurring in a document (as these are attributed by a human editor), which can be used a standard for further evaluation of the content. For WRAPIN, a corpus of MeSH-annotated web sites would have been interesting, however to create such a corpus would require major investment. The corpus used is thus limited to the MEDLINE abstract, and was further narrowed to the domain of Gastroenterology. This domain was chosen because the ARIANE modules have been validated and optimized for it. To create the corpus we wrote a detailed query for the MEDLINE database and obtained 6,768 articles.

In this set, we randomly selected 200 articles to make the evaluation, keeping only the title, the abstract and the annotated MeSH terms. The goal of the evaluation was to compare the MeSH terms automatically extracted from the abstracts by our system with those annotated by the experts in the MEDLINE citations. Our basic system is a simple lexical extractor that is described in [15, 16]. In this system, normalization is mainly supplied by the terminological resources of the MeSH (synonymous and closer expressions included). This system follows the two assumptions proposed by Cooper [12] in his "PostDoc" lexical algorithm. The first assumption is "that the medically meaningful content in free-text clinical records would be contained within noun phrases" and the second is "that all the important medical words worth recognizing in free-text noun phrases should be related to the words in the target vocabular-

[1] "Cholelithiasis", "Human", "Intraoperative care", "Retrospective studies" are MeSH terms forgotten by our system for this abstract

[2] The distinction between the major and minor MeSH term doesn't appear here

[3] 11.1% and 43.6% of precision respectively for the base system and those with UMLS

ies" (here the MeSH thesaurus). Further, our system is a regular expression-based system, which allows some insertion, deletion and substitution to treat lexical variations. It can also recognize MeSH terms on a window of five words.

Table 1: N-best MeSH terms for the basic and extended extraction

Rank	Basic extraction	Extraction with UMLS knowledge sources
1	Bile	Common bile duct calculi
2	Common bile duct	Cholecystectomy, laparoscopic
3	Bile duct	Cholangiopancreatographyendoscopic retrograde
4	Patient	Cholangiography
5	Duodenum	Cholecystectomy
6	Common bile duct calculi	Duodenum
7	Therapeutics	Common bile duct
8	Cholecystectomy	Bile duct
9	Cholecystectomy, laparoscopic	Bile
10	Cholangiography	Therapeutics
11	Cholangiopancreatography, endoscopic retrograde	Patient

Note that the evaluation method presented here is a method used normally in the information retrieval domain (TREC[1]). Usually the evaluated elements in this domain are the returned documents according to the queries. Here, in our evaluation paradigm, queries are MEDLINE abstracts, the returned documents are the MeSH terms (the candidates), and the reference is represented by the annotated MeSH term (given by the experts during the MEDLINE indexing process).

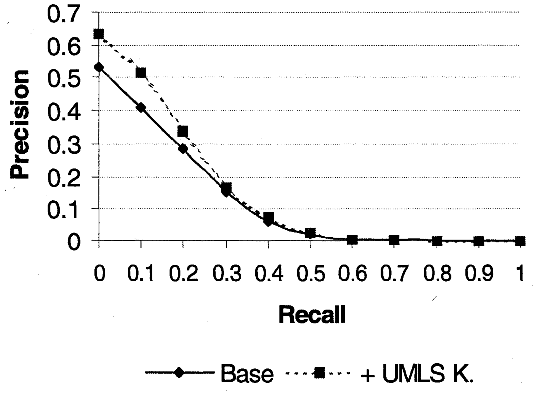

Figure 2 - Curve of comparative precision/recall with the UMLS knowledge sources.

Results

The Figure 2 shows results in terms of precision and recall (11 interpolated points for N = 20)[2] measures. For this evaluation the major and minor MeSH terms are not differentiated. Here, two systems are compared, the first noted "base" (diamond and unbroken line) is our base system and the second noted "+ UMLS K." (square and dashed line) is the basic system improved by the usage of the UMLS co-occurrences.

On the whole evaluated abstract corpus, our basic system has 53% of good propositions when it proposes only the first MeSH term while the second system has 63% of precision. The precision of the base system is 40.8% and 28.4% whereas the precision for the second is 51.6% and 33.8% respectively at 10% and 20% of recall. Further more the Figure 2 shows mainly that the integration of knowledge improves significantly the precision by 10% at the first point of recall. This improvement decreases regularly until 30% of recall.

Discussion

The WRAPIN project is an ambitious project that depends on some component efficacy. Specifically, the MeSH term extraction is a critical component because it allows identification of key concepts from a user query. This paper shows an original way of exploiting the quantitative weights of relationships that exist between concepts, as represented in the UMLS knowledge sources. Weights affected to relationships, used to refine the MeSH extraction process, are computed on the basis of the frequency of co-occurrences between major terms in the literature citations and the semantic network. These serve to measure the importance of the concomitant presence of recognized terms in a document.

Our evaluation convinced us this was a valid approach to enhance the extraction of MeSH terms from documents with the help of semantics such as those present in the UMLS knowledge sources. However, several comments are necessary regarding the evaluation we made:

- The manual annotation is made on the full articles while the automatic extraction is run on the abstracts only.
- No distinction is made between minor and major MeSH terms.
- Check tags of MEDLINE annotation are ignored by the system although they are proposed by the annotators (and so counted in the evaluation).

As a corollary to these limitations, on one hand the results are difficult to interpret in an absolute way and need further investigation, on the other hand the comparison of results obtained between several systems and the first promising outcomes should justify the interest of this work.

The precision of the MeSH term extractor in WRAPIN was successful due to its use of only the first MeSH terms encountered for its queries to the medical databases. The system is, however, not sufficient to replace a human annotator for initial indexing of a document. For this, some form of interactive process could be envisaged, both for scientific articles or for web pages.

Obviously with this first positive results we should extend our corpus for the evaluation in order to verify if this first result can

[1] Text REtrieval Conference: http://trec.nist.gov/

[2] N = 20 indicates that the automatic extraction system proposes 20 MeSH terms per abstracts. This number is largely enough when we want to query other databases

be generalized to the entire medical domain, and to other types of documents. In this evaluation, "the related terms heuristics" (scoring the result using UMLS knowledge sources) has improved the precision in the task of MeSH extraction; nevertheless it would certainly be judicious to combine this heuristics with others to obtain enhanced performance. At the moment this heuristic has been developed as a post processing system, it would also be interesting to find a way to integrate this heuristic directly during the extraction process to try to increase the recall.

Acknowledgements

The WRAPIN project, IST-2001-33260, is supported by the European Commission and the "Office Fédéral de l'Education et la Science" (OFES, Switzerland). The authors wish to thank all the WRAPIN members with their precious contribution in this ambitious project. WRAPIN is powered by Sun Microsystems.

References

[1] Boyer C, Baujard O, Baujard V, Aurel S, Selby M, Appel RD. Health On the Net automated database of health and medical information. *Int J Med Inf* 1997; 47: 27-29.

[2] Boyer C, Selby M, Appel RD. The Health On the Net Code of Conduct for medical and health Web sites; its status in 1997. *MEDNET97 - European Congress on the Internet in Medicine*, Brighton, U.K., Nov. 03 to 06, 1997. HONcode : http://www.hon.ch/HONcode/

[3] National Library of Medicine. Medical subject headings. http://www.nlm.nih.gov/mesh

[4] Joubert M, Fieschi M, Robert JJ, Volot F, Fieschi D. UMLS-based conceptual queries to biomedical information databases: an overview of the project ARIANE. *J Am Med Inform Assoc.* 1998; 5(1): 52-61.

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

[6] McCray AT, Nelson SJ. The representation of meaning in the UMLS. *Methods InfMed.* 1995; 34 (1-2): 193-201.

[7] Joubert M, Aymard S, Fieschi D, Volot F, Staccini P, Robert JJ, Fischi M. ARIANE: integration of information databases within a hospital intranet. *Int J Med Inf.* 1998; 49(3): 297-309.

[8] Masarie FE Jr, Miller RA. Medical subject headings and medical terminology: an analysis of terminology used in hospital charts. *Bull Med Libr Assoc.* 1987; 75: 89-94.

[9] 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.

[10]Brownlow ND, Hersh WR, Campbell EM. Automating concept identification in the electronic medical record: an experiment in extracting dosage information. *Proc. Annual AMIA Fall Symp.* 1996: 388-392.

[11]Zweigenbaum P, Darmoni SJ, Grabar N. The Contribution of Morphological Knowledge to French MeSH Mapping for Information Retrieval. *Proc. Annual AMIA Fall Symp. 2001*:796-800.

[12]Cooper G, Miller R. An experiment Comparing Lexical and Statistical Methods for extracting MeSH Terms from Clinical Free Text. *J Am Med Inform Assoc* 5, 1998: 62-75.

[13]Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. *Proc. Annual AMIA Fall Symp.* 2001;17-21.

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

[15]Gaudinat A, Boyer C. Automatic Extraction of MeSH terms from MEDLINEs Abstracts. *Workshop on Natural Language Processing in Biomedical Applications*, NLPBA2002: 53-57.

[16]Ruch P, Gaudinat A. Combining regular variation and nearest neighbors for efficient mapping to indexing terms. Romand 2002, Frascati. *The 2nd Workshop on Robust Methods in Analysis of Natural Language Data*, July 2002; 72-79.

Address for correspondence

Health On the Net Foundation
24 rue Michelie-du-crest,
1211 Geneva 14 Switezerland
Arnaud.Gaudinat@healthonnet.org
http://www.hon.ch/