.

Automatic Extraction of MeSH terms from Medline Abstracts

Arnaud Gaudinat, Celia Boyer
Health On the Net Foundation (HON), Geneva University Hospital, Geneva, Switzerland

Abstract:

Information retrieval has two fundamental steps, indexing and retrieval. The first step is crucial, determining the indexes which will be used during the retrieval process. One well-known approach to index a document is to use a thesaurus. In the medical field the MeSH[1] is the best one. Here we present an original evaluation based on our MeSH extractor rather than a complete information retrieval evaluation. We want to observe global performance as well as the most frequent errors. For that, we use standard tools and the CF[2] collections as a reference database. The results of this experiment are generally positive, we have obtained baseline performance data to compare with other systems and to improve our extractor. This exercise has given us new ideas to increase coverage of our system, with for example, simple NLP methods..

INTRODUCTION

Information retrieval has two fundamental steps: indexation and retrieval. The first step is crucial, as it must select the index which will be used during the retrieval process. In the medical field the existence of quality thesauruses by the NLM, such as MeSH [1], have permitted the development of a thesaurus-based retrieval system in order to guide the selection of index terms.

Our search tools – MedHunt and HONSelect [2] belong to this type of system, where the Web indexing is essentially created based on the terms of the MeSH thesaurus. In this article we primarily describe the evaluation of our extractor of MeSH terms, which is used in many levels in our retrieval systems. In this exercise we present our rationale and the methodology employed. We will also show that it is possible to use standard tools and evaluation data to evaluate an index extraction component. We will then glance through our results and draw up a list of envisaged new perspectives.

OBJECTIVES

Many of our tools depend on the selection of MeSH terms. Our HONselect system, uses the MeSH structure in the organisation of information at both the database level as well as at the user interface level. The MedHunt system (our full text search engine) is also, in part, based on the same databases. Finally, in the case of a submission of a medical web-site directly by the webmaster, we wish to offer the MeSH term extractor in order to better label the submitted medical site interactively. The choice of MeSH terms is, thus, critical to make our tools' perform optimally. The evaluation of this component of our systems is both justified and necessary.

Every system that aspires to quality must undergo an evaluation. Despite the limiting

[1] Medical Subject Headings
[2] Cystic Fibrosis Collection

framework of such evaluations, they do allow for the identification of certain 'gaps' within a system. It is necessary to use a reference index, such as the CFC, to be able to quantify improvements and to allow a valuable comparison with other systems.

In our case it is interesting to identify the MeSH terms proposed in the referee collection, created manually, and not identified automatically by our extractor tool. However the result should be balanced with the fact that the referee MeSH terms have been extracted based on the full articles of the CFC collection. This full version is not available online. Thus for the evaluation of our MeSH term extractor we had used only the abstract of the CFC collection.

The existence of a referee index may allow us to quantify the integration of every new component such as an instrument to transform terms into root terms or the use of UMLS for normalisation.

Within the European WRAPIN Project , which has as its main aim the automatic provision of reliable advice about medical sites, the extraction of index terms is a key stage.

EVALUATION METHODOLOGY

In order to realise an evaluation, it is of the utmost importance to use a referee source of information with a respectable size. Experts should index this referee source manually. The CFC (Cystic Fibrosis Collection) [3] conforms to these primary requirements in the field which interests us, i.e. the selection of MeSH terms. This database is composed of 1239 documents which were indexed using the term « Cystic Fibrosis » in the database MEDLINE of the NLM. In fact, this collection contains, amongst other things, summaries for which experts have selected a list of major MeSH terms and one of minor MeSH terms. This collection is accompanied by a database of queries (with defined results) in order to perform a complete evaluation of a document retrieval system.

Since the evaluation concerns only the extraction of MeSH terms, it was carried out according to the referee terms– the major and minor MeSH terms. However, due to the fact that these 2 referees are complementary, a third referee was added, which, in fact, regroups the first 2 types.

Within a classic framework of evaluation, we attempt to find the relevant documents relative to the end-users' requests. In our case only MeSH terms and their summaries interest us since we want to automatically find the MeSH terms for a given summary. In our evaluation paradigm, the proposed requests then become the summaries, whereas the returned documents are transformed into MeSH terms. This 'trick' allows us to use the evaluation instruments and the collection referee. Figure 1 presents the standard measure of performance [4] « Precision/Recall» of the retrieval systems within the framework of MeSH term extraction.

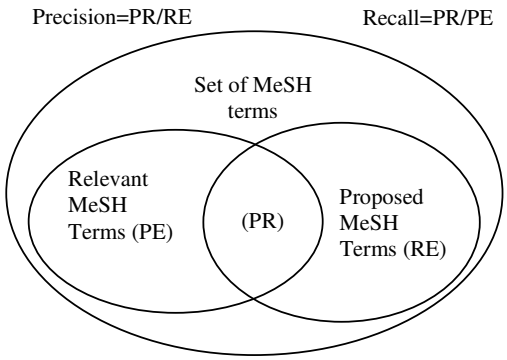

Figure 1. Precision & Recall on MeSH Term

The extraction of MeSH terms is both simple and efficient, it is achieved through an exhaustive string comparison between the full text words to be analysed and the MeSH terms from the thesaurus. This comparaison is a fuzzy match based on regular expressions allowing some insertions, substitutions and deletions, and permiting correction of misspelled words, suggesting morphologic variations. In the case of composed words, we carry out the same search as far as 5 adjacent word series. A list of stop-words permits avoidance of the selection of certain grammatical words. Finally, the MeSH terms are presented in an ordered list following diverse statistical weightings.

In order to quantify the differences between

the referee collection and the hypotheses of our extraction system, we use the program TrecEval[3] issued by TREC [6], which permits us to obtain a global representation of system performances in the form of a standard precision/recall curve, such as the one described by [4]. This program forms part of the Smart system, which takes as its parameters two files which are the referee file and the result file.

TrecEval allows us to obtain the performances for every proposed abstracts, following which, it is easy to recover the worst results and interpret them.

In order to submit the extractor hypotheses, they must be put into forms and an order corresponding to our preferred MeSH term must be indicated. We have, therefore, attributed to our MeSH terms a value of between 1 and 15 (as the extractor is configured to produce a maximum of 15 hypotheses) in the field similarity of the hypothesis file.

It is important to be able to identify the MeSH terms least well recognised in our system. With the aid of a simple comparison program, we have therefore automatically extracted both the non-pertinent MeSH terms most often proposed as well as those least well recognised by the extractor.

RESULTS

The first result is the precision/recall graph (figure 2) for the different types of subject in the CFC collection. Overall this result corresponds to that which already exists in the literature for this type of exercise. Indeed, although this evaluation is difficult to compare with the one made by de Bruinj [11] and the other by Franz [12] (which differs in term of collection, thesaurus and system), the first system gives a precision of 65.2% for the first proposed term and the second present a precision of 50.4% for its best score. For the minor subjects and the 2 types of subject, the precision measure has a tendency to fall rapidly once there are multiple hypotheses. As for the results on the major subjects, these are very good, especially if we consider that they are in essence, the most important. Concerning the latter, considering that the system proposes 15 MeSH terms, we achieve a precision close to 51% for the first 4 terms proposed, with an important precision loss around the eighth MeSH term.

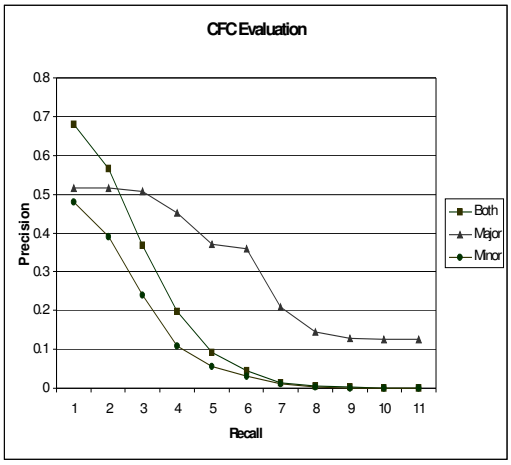

Figure 2. Precision/Recall Curve

Even if the abstract lists for which no MeSH term was proposed by the system underlines some of the gaps in our system, the errors result mainly from the difference between terms manually selected by experts and the real content of the full article within the collection. This difference is not easily measured and does not allow us to achieve conclusions about absolute values of the evaluation.

The list of non-relevant words proposed by the extractor is shown in table 1. We can find very general terms in the medical field such as « PATIENT » or « DISEASE » and others terms very specific to the collection such as « FIBROSIS ». These are, in fact, terms which do not have a discriminatory relevance in this collection. These general terms could be penalized using the « idf factor » (inverse document frequency) calculated on the 20000 terms of the Mesh thesaurus. This must, nevertheless, be tested.

MeSH term	No. of occurrences
CHILD	98
SOCIAL-VALUES	101
DIAGNOSIS	143

[3] ftp://ftp.cs.cornell.edu/pub/smart/

THERAPEUTICS	149
CELLS	157
ATTENTION	158
CALCIUM-SULFATE	243
DISEASE	389
PATIENTS	707
FIBROSIS	957

Table 1. False positive MeSH terms

Table 2 gives us an indication of the MeSH terms that are never proposed by the system - even if they are present in the referee database. The MeSH term « HUMAN » is present in many documents, of which there are 1205 abstracts where the extractor did not propose it. Observing these terms, we are able to regroup most of them into 2 categories, that of age-group and that of gender and species.

MeSH term	Number of occurrences
MIDDLE-AGE	47
…	…
INFANT-NEWBORN	141
SUPPORT-U-S-GOVT-P-H-S	178
INFANT	226
ADULT	327
CHILD	348
CHILD-PRESCHOOL	361
ADOLESCENCE	404
CYSTIC-FIBROSIS	439
MALE	465
FEMALE	469
HUMAN	1205

Table 2. False negative MeSH terms

In fact, regarding the age groups, we find (the percentage is the relationship between the occurrence of the term and the number of summaries):

- MIDDLE-AGE (3.8%)
- INFANT-NEWBORN (11.4%)
- INFANT (18.2%)
- ADULT (26.4%)
- CHILD (28.1%)
- CHILD-PRESCHOOL (29.2%)
- ADOLESCENCE (32.6%)

In our evaluation 32,6% of the documents contain the MeSH term « ADOLESCENCE ", which is not recognized.

This problem is due to the complexity of language, because a large number of ways to express the age of a person in natural language exists. For example the MeSH term « CHILD » should be selected if we find the series of words « 12 years old » in the text, while the series of words « the age of 6 month to 1 year » should select the MeSH term « INFANT ».

Consequently, the use of a micro-grammar describing all the possible ways to express an age-group in order to select all the MeSH terms presented in table 3 should solve this problem.

MeSH Term	Age
NEWBORN	Birth to1 month
INFANT	1 to 23 months
PRESCHOOL	2 to 5 years
CHILD	6 to 12 years
ADOLESCENCE	13 to 18 years
ADULT	19 to 44 years
MIDDLE AGE	45 to 64 years
AGED	65 to 79 years
80 AND OVER	More than 79 years

Table 3. MeSH terms for age groups

CONCLUSION

The use of the CFC allows us to perform a realistic evaluation of our MeSH term extraction on both qualitative and quantitative

levels. Nevertheless, the results need to be considered as relative while examining the quality of the MeSH terms chosen by the experts in the CFC, which is based on the full document and not on the abstract. The main goal is, nonetheless, achieved since we now have at our disposition a solid comparison base for future enhancements.

We have shown that it is possible to realize an evaluation of one of the principal components of a retrieval system, which is the selection of indexes and this with classic evaluation instruments and data.

This evaluation allows us to give weight to a component, which is often put aside in search engines, namely the linguistic component. In our case a simple regular grammar should allow for the selection of the correct MeSH term in the age-group category.

The next stages will be to use the CFC collection in order to evaluate, as a whole, the quality of the search engine (indexation and search) of HONselect[4] and MedHunt[5] and also to perform the same types of evaluations on another collection, such as OHSUMED [11], which is larger in size and is used in TREC.

In fact, we are convinced that it is only from a large evaluation base that improvements to our system can be quantified in a more realistic manner even if theses collections are fare from the Web.

The MeSH extraction is very closely related to the relevancy of the search results of internal or external sources. The enhancement of the MeSH term extraction will have a key role into the WRAPIN European project.

ACKNOWLEDGEMENTS/THANKS

The WRAPIN (Worldwide online Reliable Advice to Patients and Individuals) 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 the following for their discussions and valuable comments for this study: Patrick Ruch from the DIM (Medical Informatics Division, Geneva University Hospital, Geneva, Switzerland), the reviewers for their relevant suggestions, Vincent Baujard for his advice and cooperation and Ken Dobruskin for help with proofreading.

[4] A multilingual and intelligent search tool integrating heterogeneous web resources: http://www.hon.ch/HONselect/Browse.html

[5] A full text information retrieval system: http://www.hon.cn/MedHunt

References

1. Lowe H, Barnet G. Understanding and using the medical subject headings (MeSH) vocabulary to perform literature searches. Journal of the American Medical Association. 1994;271:1103-1108.
2. Boyer C, Baujard V, Griesser V, Scherrer JR. HONselect: a multilingual and intelligent search tool integrating heterogeneous web resources. International Journal of Medical Informatics. 2001;64:253-258.
3. Shaw WM, Wood JB, Wood RE, Tibbo HR. The Cystic Fibrosis Database: Content and Research Opportunities. LISR 1991;13:347-366.
4. Baeza-Yates R, Ribeiro-Neto B. Retrieval performance evaluation. In: Modern Information Retrieval. New York: ACM Press,1999;73-97.
5. Baeza-Yates R, Ribeiro-Neto B. MODELING: Classic Information Retrieval. In: Modern Information Retrieval. New York: ACM Press,1999; 25-33.
6. Harman D. Overview of the second Text Retrieval Converence (TREC-2). Information Processing and Management. 1995;31;271-289.
7. Hersh WR, al. OSHUMED: an interactive retrieval evaluation and new large test collection for research,. In Proceedings of the 17th Annual International ACM Special.
8. Salton G, McGill MJ. Introduction to modern information retrieval. New York: McGraw-Hill Book.1983.
9. Aronson AR, Rindflesch TC, Browne AC. Exploiting a large thesaurus for information retrieval.Proceedings of RIAO. 1994;197-216.

10. Aronson AR. Effective Mapping of Biomedical Text to the UMLS Metathesaurus: The MetaMap Program. Proc AMIA Symp. 2001;17-21.
11. de Bruijn LM, Verheijen E, van Nes FL, and Arends JW: Assigning SNOMED codes to natural language pathology reports. In: J. Brender et al. (eds): Medical Informatics Europe, Copenhagen 1996;198-202.
12. Franz P, Zaiss A, Schulz S, Hahn Udo et Klar R. Automated coding of diagnoses--three methods compared. Proc AMIA Symp. 2000;250-4.
13. Nadkarni PM, Chen RS, Brandt CA. UMLS Concept Indexing for Production Databases: A Feasibility Study**.** Journal of American Medical Informatics Association, 2001;8:80-91.
14. Hersh WR. Information Retrieval: A Health Care Perspective.New York: Springer-Verlag,1996.
15. Hersh WR, Hickam D. A comparison of retrieval effectiveness for three methods of indexing medical literature. AmJ Med Sci. 1992;303:292-300.
16. Hersh WR, Hickam DH, Leone TJ, Words, concepts, or both: optimal indexing units for automated information retrieval, Proceedings of the 16th Annual Symposium on Computer Applications in Medical Care. 1992:644-648.