A Benchmark Evaluation of the French MeSH Indexers

A. Névéol[1,2], V. Mary[3], A. Gaudinat[4], C. Boyer[4], A. Rogozan[1], and SJ. Darmoni[1,2]

[1] PSI Laboratory, Rouen, France - {aneveol ,arogozan}@insa-rouen.fr}
[2] CISMeF & CGIS, Rouen, France - stefan.darmoni@chu-rouen.fr
[3]Rennes Medical School, France - vincent.mary@univ-rennes1.fr
[4]HON Foundation, Geneva, Switzerland - Firstname.Lastname@healthonthenet.org

Abstract. The increasing demand on both practitioners and librarians to encode medical documents with controlled vocabularies calls for automatic tools and methods to help them perform this task efficiently. This paper presents the Benchmark evaluation of the French MeSH indexing systems carried out under the umbrella of the VUMeF consortium. The CISMeF, NOMINDEX and HONMeSHMapper systems are introduced, and evaluated on a set of 82 resources randomly taken from the CISMeF catalogue. The automatic MeSH indexing produced by each system was compared to the manual gold standard provided by the CISMeF medical librarian team. The automatic systems achieve at best a precision close to 50% at rank 1 (HONMeSHMapper, CISMeF) and HONMeSHMapper achieves the best overall F-measure. A qualitative evaluation of the indexing provided indicates that all systems tend to misevaluate the specificity of the terms to retrieve.

1 Introduction

The internet has become a very prosperous source of information in numerous fields, including health and molecular biology. Several projects have been initiated in order to meet the users' information needs related to these fields. Among them, the Health On the Net foundation (HON[1]) aims at guiding both lay and specialist audiences to trustworthy medical information. HON has developed automatic search engines to crawl and index the web, and an accreditation system based on their HONcode principles. Some 4,600 websites are currently accredited and annually reviewed. The Nomindex project[2] aims at organising health information for a more efficient retrieval. CISMeF[3] (French acronym of Catalogue and Index of Medical On-Line Resources) describes and indexes more than 14,000 resources of institutional health information in French. Indexing is a decisive step for the efficiency of information retrieval within these systems, and if performed manually, it is a highly time consuming task. Automatic tools have been developed for MeSH indexing in English as early as the 80s [1]. More recently, such tools have also been available for French. This paper presents the results of the Benchmark evaluation of the French

[1] http://www.hon.ch/ (accessed on February 1st, 2005)
[2] http://www.med.univ-rennes1.fr/nomindex/ (accessed on February 1st, 2005)
[3] http://www.cismef.org (accessed on February 1st, 2005)

MeSH indexing systems which was carried out in 2004 under the umbrella of the VUMeF [2] consortium. The aim of this evaluation is twofold: first, it provides a comparison of the systems. Secondly, through the analysis of the results, the strengths and weaknesses of each system may be identified. If appropriate, the complementarity of the systems – or of the resources they use, may be exploited.

2 Material and Methods

2.1 The French MeSH Indexing Systems

CISMeF - Natural Language Processing Approach (NLP)

This approach (detailed in [3]) is built on the three-step manual indexing procedure: analysis of the resource to be indexed, translation of the emerging concepts into the appropriate controlled vocabulary (here, the MeSH) and revision of the resulting index. First, a MeSH dictionary is used to extract medical concepts. The variants of the concepts (inflected forms, synonyms, etc.) are taken into account to compute the frequency of each concept. The dictionary contains the necessary information to translate the concepts into MeSH terms. A tf*idf normalization is then used to compute relevance scores for each MeSH term. The hierarchical information drawn from the MeSH is used to select and promote the most precise terms. Moreover, recurring check tags are promoted at the top of the candidate list to ensure their selection. Eventually, indexing rules are applied in order to revise the candidate list before the final index selection using a breakage function [3]. Although this system is able to retrieve isolated keywords, it was conceived to retrieve keyword/qualifier pairs.

NOMINDEX

Nomindex [4] was developped in order to identify medical concepts in natural language sentences. Then, these concepts are stored in a database which may be used for information retrieval. Nomindex uses a lexicon derrived from the ADM [5] (Assisted Medical Diagnosis) knowledge base which contains 130.000 terms, including associated words, compound words, prefixes and suffixes. First, document words are mapped to ADM terms and reduced to reference words (for instance, "cephalalgia" is mapped to "headache"). Then, ADM terms are mapped to the equivalent French MeSH terms, and also to their UMLS Concept Unique Identifier. Finally, every reference word of the document is then attributed its corresponding UMLS CUI. A relevance score (tf*idf) computed for each concept found in the document, is used in various tools : keyword identification, document similarity and automatic document synthesis.

HONMeSHMapper

HONMeSHMapper was developed in 1997 for the automatic categorization and retrieval of online medical documents. It is encapsulated in a more generic term extractor which uses generic terminological resources such as the UMLS. Initially developed for HONselect and enhanced through the years, HONMeSHMapper has become a major component of the WRAPIN project [6] for the task of MeSH

keyword extraction and mapping. Initially, it was a lexical mapper [7] based on Cooper's assumptions [8] that (1) "the medically meaningful content in free-text clinical records would be contained within noun phrases" and (2) "all the important medical words worth recognizing in free-text noun phrases should be related to the words in the target vocabularies" (here, the MeSH thesaurus). In this system, normalization is supplied by MeSH terminological resources (including synonyms and close expressions) and by a stemmer. HONMeSHMapper is a regular expression-based system which can also recognize compound MeSH terms within a window of five words. A bag-of-words approach takes into account the distribution of components of compound MeSH terms found in the full text. Finally, a weight is assigned to each MeSH term retrieved according to (1) the inverse frequency of the term in the document and (2) its position within the MeSH hierarchy.

2.2 Evaluation corpus and measures

The corpus used for this evaluation is composed of 82 resources randomly selected in the CISMeF catalogue. It contains about 235,000 words (1.7 Mb.). The corpus has been indexed by five professional indexers. In the literature [9], the manual indexing is considered as a gold standard to which the automatic indexing may be compared, although the inter-expert variability is high [10]. The average number of isolated keywords manually assigned to a resource in the evaluation corpus is 7.56 +/- 6.92.

The evaluation measures used are precision and recall. We also used the F-measure, which combines them with an equal weight. In the gold standard indexing used as a reference, the indexing terms consist of MeSH keyword/qualifier *pairs*. However, two of the indexing systems (NOMINDEX and HONMeSHMapper) retrieve isolated keywords. Therefore, we have focused the evaluation on the retrieval of keywords. We have considered that retrieving an isolated keyword, where the gold standard advocates the same keyword associated to a qualifier, was correct. For example, if <*hepatitis*> was retrieved where <*hepatitis/therapy*> was expected, we considered that the index term had been correctly retrieved.

3 Results

Table 1 shows the precision and recall (P-R) obtained by each system.

Rk	NOMINDEX	HONMeSHMapper	CISMeF- TAL -
	P – R	P – R	P – R
1	13.25 - 2.37	45.78 - 8.63	45.78 - 7.42
4	12.65 - 9.20	31.93 - 26.41	30.72 - 22.05
10	12.53 - 22.55	20.61 - 36.96	21.23 - 37.26
50	6.20 - 51.44	7.76 - 57.81	7.04 - 48.50

Table 1: Precision and recall of each system at fixed ranks

Figure 1 allows a comparison of the three systems through F-measure:

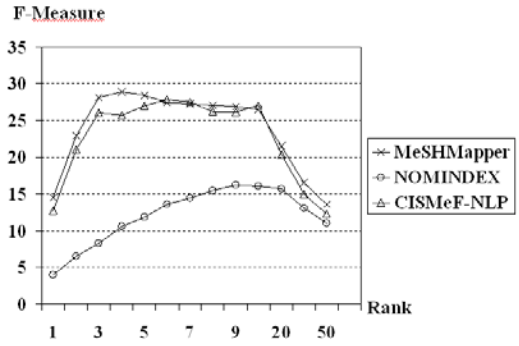

Figure 1: plot of F-Measure vs. fixed ranks for each indexing system.

4 Discussion

Global performances of the systems

According to Table 1, the systems achieve at best a precision of 45% at rank 1 (HONMeSHMapper, CISMeF). HONMeSHMapper and CISMeF show a similar precision at all ranks, but the recall is higher for HONMeSHMapper. Figure 1 reflects this observation, as HONMeSHMapper achieves the best overall F-measure. MTI (Medical Text Indexer) [9] obtained a precision of 29% and a recall of 55% at rank 25 on a corpus of 273 articles. To compare the experiments, we may observe that (1) the evaluation corpus was different (scientific articles *vs.* web pages) and (2) the terminological resources or English are more comprehensive than those available in French: in 2005, about 50.000 MeSH synonyms remain to be translated into French.

Qualitative Analysis

A qualitative analysis of the terms retrieved by each system shows that the "noise" does not result from the retrieval of irrelevant terms. Most of the terms retrieved that are not selected by the human indexers are in fact either too broad (the indication on the resource content is too vague to be useful to the users) or too narrow (the concept referred is not sufficiently developed in the resource, so that users would not be satisfied with the information provided). Deciding whether the degree of specificity of each term retrieved is adequate would improve the performance of the three systems.

Perspectives

The terminological resources used by all three systems have different origins, and may be complementary. A previous evaluation of NOMINDEX [11] showed that specific updates of the lexicon could improve the system performance. Therefore, sharing resources may benefit all systems.

A recent evaluation of the American MeSH indexing system MTI [9] showed the advantage of combining different approaches (NLP & statistical methods) and filtering rules. CISMeF is currently testing the combination of the NLP system described with a statistical (k-NN) approach for keyword/qualifier pair indexing [12]. The combined approach was also evaluated on the same corpus for pair retrieval.

Although the task was more difficult, the performances obtained matched those of isolated keyword retrieval. Therefore, the system resulting from the combination of NLP and statistical approaches for keyword/qualifier retrieval will be used for indexing resources to be added to the CISMeF catalogue in a semi-automatic mode. For HONMeSHMapper, the use of CISMeF manually indexed resources will allow the development of a knowledge-based approach, complementing the lexical approach already in use to suggest 5 keywords (WRAPPIN).

5 Conclusion

This paper presents a comparative evaluation of three MeSH indexing systems for French. MeSH isolated keywords were retrieved by CISMeF, HONMeSHMapper and NOMINDEX from the 82 resources of the evaluation corpus and compared to the manual gold standard. The best precision (45%) is achieved by HONMeSHMapper and CISMeF at rank 1. HONMeSHMapper shows the best overall F-measure. Sharing lexical resources used by all systems could enhance the performance. A qualitative evaluation of the indexing indicated that all systems could also be improved by judging more accurately the specificity of the terms to retrieve.

References

1. Humphrey SM., and Miller NE. Knowledge-based indexing of the medical literature: The Indexing Aid Project. J Am Soc Inf Sci, 38(3):184-96. (1987)
2. Darmoni, S.J. and Consortium VUMeF. VUMeF: extending the French involvement in the UMLS Metathesaurus. AMIA Annu Symp Proc. 2003;:824. (2003)
3. Néveol, A., Rogozan, A., Darmoni, S.J. : Automatic indexing of online health resources for a French quality controlled gateway. In IP & M, in press. (2005).
4. Pouliquen, B., Delamarre, D., Le Beux, P., Indexation de textes médicaux par extraction de concepts et ses utilisations, JADT'2002, St Malo, France, March 2002; (2) 617-628
5. Lenoir P., Michel JR., Frangeul C., and Chales G, Réalisation, développement et maintenance de la base de données A.D.M. Médecine informatique. 1981; 6 51--6.
6. 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. In Medinfo. 2004;2004:356-60.
7. Gaudinat A, Boyer C. Automatic Extraction of MeSH terms from MEDLINEs Abstracts. Workshop on Natural Language Processing in Biomedical Applications, 2002: 53-57.
8. 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.
9. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004;2004:268-72.
10. Funk ME., Reid CA. and Mc Googan LS. Indexing consistency in MEDLINE. Bull. Med. Libr. Assoc. 71(2):176-183. (1983).
11. Mary V, Pouliquen B, Le Duff F, Darmoni SJ, Segui A, Le Beux P. Automatic conceptual indexing of French pharmaceutical theses. Stud Health Technol Inform. 2002;90:388-92.
12. Névéol A., Rogozan A., Darmoni SJ. Indexation automatique de ressources de santé à l'aide paires de descripteurs MeSH. TALN 2005. (in press).