A Benchmark Evaluation of the French MeSH Indexing Systems

Aurélie Névéol[1,2], Vincent Mary[3], Arnaud Gaudinat[4],
Alexandrina Rogozan[1], and Stefan J. Darmoni[1,2]

[1] PSI Laboratory - FRE 2645 CNRS - INSA de Rouen, France
{aneveol ,arogozan}@insa-rouen.fr}
[2] CISMeF & L@stics - Rouen University Hospital and Rouen Medical School, France
stefan.darmoni@chu-rouen.fr http://www.cismef.org
[3]Rennes Medical School, France
vincent.mary@univ-rennes1.fr
[4]HON Foundation, Geneva, Switzerland
Arnaud.Gaudinat@healthonthenet.org

Abstract. The increasing number of health documents available in electronic form, and the demand on both practitioners and librarians to encode these 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 for a sample resource indicates that all systems tend to misevaluate the specificity of the terms to retrieve.

1 Introduction

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' need to find precisely what they are looking for among the numerous documents related to these fields available online. Among them, the Health On the Net foundation (HON[1]) aims at guiding both lay and specialist audiences to trustworthy medical information in various European languages. 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. All the pages belonging to these sites are indexed. Specialised search engines developed for the medical field can now supply trustworthiness

[1] http://www.hon.ch/ (accessed on February 1st, 2005)

indication based on HONcode accreditation for each website referenced with HON. The Nomindex project[2] was also initiated in order to organise health electronic information for a more efficient retrieval. CISMeF[3] (French acronym of Catalogue and Index of Medical On-Line Resources) describes and indexes the most important resources of institutional health information in French [1]. It currently contains more than 14,000 resources, and is updated manually with 55 new resources each week. Indexing is a decisive step for the efficiency of information retrieval within the CISMeF catalogue, and it is also one of the most time consuming tasks for the librarians.

This observation shows that it is necessary to develop automatic tools to assist the human indexers in their work. Such systems have been developed for MeSH indexing in English as early as the 80s [2]. More recently, MeSH indexing tools have also been available for French. This paper presents the results of the Benchmark evaluation of the French MeSH indexing systems which was carried out in 2004 under the umbrella of the VUMeF [3] consortium[4]. VUMeF aims to enrich the terminological resources available for the biomedical domain in French, and specifically focuses on the translation of thesauri already included in the meta-thesaurus UMLS (e.g. MeSH, SNOMED). The consortium is also concerned with the development of new thesauri (e.g. CCAM), and with the evaluation of the impact of these resources on the software tools exploiting them - in particular, indexing systems.

2 Material and Methods

This section introduces the different elements involved in the evaluation, viz. the French MeSH indexing systems developed both in France and Switzerland, the evaluation corpus and evaluation methods.

2.1 The French MeSH Indexing Systems

CISMeF - Natural Language Processing Approach (NLP)

This approach (detailed in [4]) 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. As recommanded by [7], 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

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

[3] http://www.cismef.org (accessed on February 1st, 2005)

[4] http://www.vidal.fr/vumef/ (last visited on February 1st, 2005)

list to ensure their selection. Eventually, indexing rules are applied in order to revise the candidate list before the final index selection using the breakage function described below. Although this system is able to retrieve isolated keywords, it was conceived to retrieve keyword/qualifier pairs. This latter configuration will be used as a (semi)automatic indexing tool in the CISMeF indexing process.

NOMINDEX

The purpose of Nomindex [5] is to recognize concepts in a sentence written in natural language and to use them to create a database allowing to search documents. Nomindex uses a lexicon derrived from the ADM [6] (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 [7]), computed for each concept found in the document, is used in various tools : keyword identification, document similarity and automatic document synthesis.

HONMeSHMapper

The HONMeSHMapper system was developed in 1997 along with MARVIN (Multi-Agent Retrieval Vagabond on Information Networks) in order to automatically retrieve and categorise online medical documents. These projects were supported by the Swiss National Fund for Scientific Research-robot .

HONMeSHMapper is encapsulated in a more generic term extractor which is able to deal with other terminological resources such as UMLS, but which has also been used successfully with a specific medical ontology. Initially developed for HONselect and enhanced through the years, it has become a major component of the WRAPIN project [9] for the task of keyword extraction and keyword mapping. Available in 7 languages, it addresses the problem of Information Retrieval on the Web. Initially, It was a lexical mapper as described in [10]. This system follows the two assumptions proposed by Cooper [11] 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 vocabularies" (here, the MeSH thesaurus). In this system, normalization is mainly supplied by the terminological resources of the MeSH (synonymous and closer expressions included), but also by a stemmer (such as Porter). The HONMeSHMapper system is a regular expression-based system which can also recognize compound MeSH terms within a window of five words. A bag-of-words approach is also used to take into account the distribution of components of compound MeSH terms found in the full text. Finally, a first weight is assigned according to the inverse frequency of the MeSH term in our indexed Web page (from MedHunt and HONcodeHunt). A second weight is computed according to the different MeSH hierarchical classes.

Breakage Function

Let *N* be the number of indexing candidate keywords (or pairs) retrieved with one of the methods described above. Let S_i be the score assigned to the *i*-th candidate. Let us assume that the candidates are ordered by decreasing scores, so that $S_1 > \ldots > S_i > \ldots > S_N$. For *i* =1, …, *N*-1, we compute $F = \frac{\mathrm{S_i - S_{i+1}}}{\mathrm{S_i + S_{i+1}}}$. The final index threshold is *i* such that F reaches a maximum. The purpose of this breakage function is to select an *adaptive* threshold for each resource indexed rather than arbitrarily retaining a fixed number of candidates. Selecting a different-size index for each resource reflects both the practise of human indexers and the fact that an automatic system may not be equally efficient on every resource.

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 altogether, which represents about 1.7 Mb. These resources have been manually indexed by five professional indexers in the CISMeF team. In the literature [12], the manual indexing is considered as a gold standard to which the automatic indexing produced by each system is compared, although the inter-expert variability is high [13]. The average number of isolated keywords used by the indexers to index a resource in the evaluation corpus is 7.56 +/- 6.92. The average number of keywords or keyword/qualifier pairs used to index a resource in the evaluation corpus is 10.74 +/- 9.80.

The evaluation measures used are precision and recall. For a better comparison of the systems, we also used the F-measure, which combines both precision and recall with an equal weight [14]. More specifically, precision corresponds to the number of indexing terms properly retrieved over the total number of terms retrieved. Recall corresponds to the number of indexing terms properly retrieved over the total number of terms expected. In the gold standard (manual) 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 <*diabetes mellitus*> was retrieved where <*diabetes mellitus/drug therapy*> was expected, we considered that the index term had been correctly retrieved. Similarly, if <*diabetes mellitus/drug therapy*> and <*diabetes mellitus/prevention & control*> were expected according to the gold standard, we considered that the automatic systems should retrieve the keyword <*diabetes mellitus*>.

The indexing of a specific resource was also analysed by an indexing expert (BT) and the keywords (or pairs) retrieved by each system (other than those appearing in the gold standard) were classified as: "irrelevant" (IR), "too broad" (TB), "too precise" (TP), or "relevant" (RE).

3 Results

Table 1 shows the precision and recall (P-R) obtained by each system. We have also used the breakage function described in 2.1. The last line of Table 1 shows the average precision and recall at the threshold and the average threshold (between brackets).

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
T	**9.70 - 11** **(T=6,6)**	**42.23 - 19.80** **(T=4.6)**	**29.93 - 29.11** **(T=12)**

Table 1: Precision and recall of each system at fixed ranks, and adaptive threshold.

Figure 1 allows a comparison of the three systems through F-measure. We can see that the F-measure increases steadily until rank 10 for NOMINDEX. For HONMeSHMapper and CISMeF, the F-measure increases until rank three, and remains stable until rank 10.

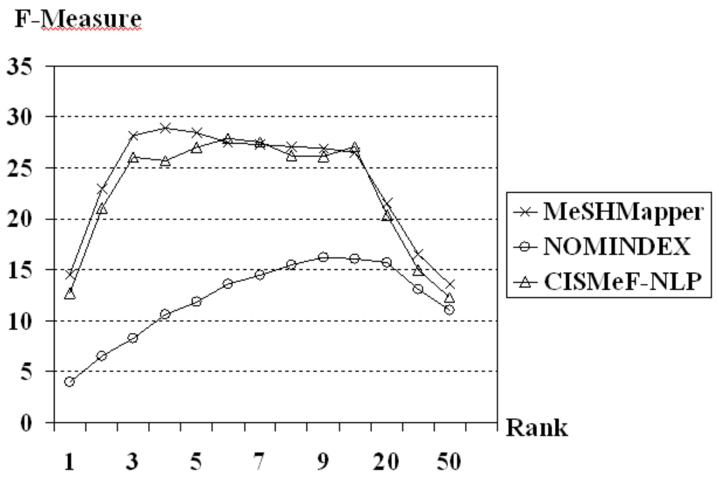

Figure 1: plot of F-Measure vs. fixed ranks for each indexing system.

Table 2 presents the fifteen first candidates retrieved by each system for the indexing of a sample corpus resource. As an indication, we also give the fifteen first candidates retrieved by the CISMeF indexing system for keyword/qualifier extraction[6].

NOMINDEX	**HONMeSHMapper**
pathological conditions, signs and symptoms *(TB)*	**irritable bowel syndrome** *(TN)*
gastrointestinal diseases *(TB)*	**diarrhea**
diarrhea	**inflammatory bowel diseases***(TN)*
signs and symptoms, digestive *(RE)*	**acute disease** *(IR)*
pathologic processes *(IR)*	**intestinal diseases** *(TB)*
digestive system diseases *(TB)*	intestine, small*(TN)*
intestinal diseases *(TB)*	tropical medicine *(RE)*
signs and symptoms*(TB)*	infection *(TB)*
disease *(TB)*	receptors, proteinase-activated *(IR)*
bacterial infections and mycoses*(IR)*	bacterial infections *(TB)*
lactose*(TN)*	travel
infection*(TB)*	sprue, tropical *(TN)*
travel	cyclosporiasis *(TN)*
irritable bowel syndrome*(TN)*	colonic diseases *(TB)*
malabsorption syndrome *(TN)*	lactose intolerance *(TN)*
CISMeF – NLP[5]	**CISMeF – MeSH *pairs* extractor[6]**
Diarrhea	**Travel**
travel	**tropical medecine** *(RE)*
syndrome*(TB)*	diarrhea/etiology
signs and symptoms *(TB)*	diarrhea/diagnosis *(RE)*
colon *(TB)*	diarrhea
lactose *(TN)*	pregnancy*(IR)*
health *(TB)*	infant*(IR)*
enteritis*(TN)*	syndrom*(TB)*
dietary fiber*(TN)*	colon*(TN)*
colonoscopy *(TN)*	infant diarrhea/therapy*(IR)*
continuity of patient care*(IR)*	water purification*(TN)*
intestin, small*(TN)*	international cooperation *(IR)*
canada*(IR)*	France*(IR)*
amebiasis*(TN)*	Infant diarrhea*(IR)*
weight loss *(TN)*	pediatrics/education*(IR)*

Table 2: Automatic indexing proposed by each system for a sample resource (http://www.phac-aspc.gc.ca/publicat/ccdr-rmtc/98vol24/24sup/dcc1.html - accessed on 01/02/05)

The keywords (or pairs) above the adaptive threshold are shown in bold characters. The relevant index terms (i.e. selected by a human indexer in the gold standard indexing) are underlined. A total of four terms were expected according to the gold

[5] The system was used to retrieve isolated keywords.

[6] This indexing results from the combination of the CISMeF-NLP system used in pair retrieval mode with a statistical system based on the k-nearest neighbours approach also retrieving MeSH pairs.

standard. <*diarrhea*>, <*travel*>, <*diarrhea/etiology*>, and <*diarrhea/therapy*>. For isolated keyword retrieval, NOMINDEX, HONMeSHMapper and CISMeF-NLP were expected to retrieve <*diarrhea*> and <*travel*>. For pair retrieval, all four terms were expected.

4 Discussion

Global performances of the systems

According to Table 1, The automatic systems achieve at best a precision of 45% at rank 1 (HONMeSHMapper, CISMeF-NLP). HONMeSHMapper and CISMeF-NLP 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. A detailed analysis of the results will allow us to identify weaknesses in each system, and the overall results will help us validate the improvements.

The automated use of HONMeSHMapper to suggest 5 MeSH terms seems reasonable. It is currently used in WRAPIN for indexing and query analysis, where its multilingual capabilities are used to transform user queries into 5 or more languages. This system is also used to assist reviewers with categorization during the accreditation process.

Aronson et al. obtained a precision of 0.29 and a recall of 0.55 at rank 25 for the MTI (Medical Text Indexer) System [12] for an indexing task of 273 articles. The difference in results can be explained by two main factors:

1.) The resources used were different (scientific articles vs. web pages; MeSH terms were probably more numerous in the scientific articles), and
2.) The English-language terminological resources are more comprehensive than those available in French, especially considering the use of UMLS by MTI. In the 2005 version of the MeSH, more than 50.000 American synonyms remain to be translated into French.

Relevance of the Threshold

The threshold function is efficient for HONMeSHMapper and CISMeF in terms of maximizing the precision. The precision at the threshold is comparatively higher than the precision at the equivalent fixed rank (eg. 42% at threshold 4.6 vs 20% at rank 5 for HONMeSHMapper). For NOMINDEX however, the threshold function is not efficient (the precision at threshold 6.6 is 9,7 % vs. 13% at rank 7). For CISMeF, the F-measure at the Threshold is actually superior to the F-Measure at any given fixed rank (F-measure at threshold is 29,51 vs. 27.89 – max at rank 6). For HONMeSHMapper, it is not the case (F-measure at threshold is 26,9 vs. 28,91 – max at rank 4). Moreover, for HONMeSHMapper, the F-measure is quite stable between ranks 3 and 10, so the Threshold function does not properly select the rank where it reaches a maximum.

Analysis of Table 2

Table 2 shows the first fifteen terms retrieved by each system for one sample resource of the evaluation corpus. For this particular resource, the terms retrieved by

NOMINDEX were considered "too broad" or "irrelevant", except for <*signs and symptoms, digestive*> which was "relevant". In fact, we can observe that NOMINDEX retrieves several keywords belonging to the same MeSH hierarchy, such as <*pathological conditions, signs and symptoms*>, <*signs and symptoms, digestive*>, <*signs and symptoms*> and <*diarrhea*>.

The terms retrieved by HONMeSHMapper were equally "too broad", "too precise" or "irrelevant" except for <*tropical medicine*> which was "relevant".

The terms retrieved by CISMeF-NLP were equally "too broad" or "too precise, except for <*diarrhea/diagnosis*> which was "relevant".

The *pairs* retrieved by CISMeF were mostly "irrelevant" except for <*diarrhea/diagnosis*> and <*tropical medicine*> which were "relevant". It is interesting to note that some of the irrelevant terms, such as <*infant diarrhea/diagnosis*> or <*infant diarrhea*> could be easily corrected by a human reviewer.

After this review, two terms considered relevant (<*signs and symptoms, digestive*>, <*diarrhea/diagnosis*> and <*tropical medicine*>) were added to the indexing of this resource. This example highlights the inter-expert variability and the important role of the super-indexer, the chief medical librarian in charge of checking the manual indexing of other medical librarians.

This analysis shows that the "noise" of the systems 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 to by the term is not sufficiently developed in the resource, so that users would be disappointed with the material they were looking for in relation with said concept).

Deciding whether the degree of specificity of each term retrieved is adequate would greatly improve the performance of the three systems that were evaluated.

Perspectives

The terminological resources used by all three systems have different origins (CISMeF, ADM and WRAPIN), and may be complementary. They could be used to enrich the French Specialist Lexicon developed in the UMLF project [15]. These resources could then be shared by all systems, in order to increase their performance. In the case of NOMINDEX, a previous evaluation [8] demonstrated that the system performance could be improved by specific updates of the lexicon.

A recent evaluation of the American MeSH indexing system MTI [12] showed the advantage of combining different approaches (NLP & statistical methods) and several filtering rules. CISMeF is currently testing the combination of the NLP system described with a statistical (k-NN) approach for keyword/qualifier pair indexing [16]. The combined approach also evaluated on the same corpus for pair retrieval. Although the task was more difficult, the performances obtained were similar to those of CISMeF-NLP and HONMeSHMapper for 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. Resources on topics widely covered in CISMeF may be indexed automatically, and for other resources, the automatic indexing proposed by the system shall be reviewed by human indexers.

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. This will certainly lead to improvement in the results. The use of a lexicon containing the most familiar medical terms would also be an advantage, considering that web pages are generally less technical than scientific articles.

5 Conclusion

This paper presents a comparative evaluation of three different MeSH indexing systems for French. MeSH isolated keywords were retrieved by CISMeF-NLP, 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-NLP at rank 1. HONMeSHMapper shows the best overall F-measure. Sharing lexical resources used by all systems could enhance the performances. Moreover, a qualitative evaluation of the indexing provided for a sample resource indicated that all systems could be improved by judging more accurately the specificity of the terms to retrieve.

Acknowledgments

The authors would like to thank Benoît Thirion for his insight in the analysis of the indexing results.

References

1. Darmoni, S.J., Leroy, J.P., Baudic, F., Douyère M., Piot, J., Thirion, B. : CISMeF: a structured health resource guide, in Methods of Information in Medicine, 39(1):30-35 (2000).
2. 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)
3. Darmoni, S.J., Jarousse E., Zweigenbaum P., Le Beux P., Namer F., Baud R., Joubert M., Vallée H., Cote RA., Buemi A., Bourigault D., Recourcé G., Jeanneau S., Rodrigues JM. VUMeF: extending the French involvement in the UMLS Metathesaurus. AMIA Annu Symp Proc. 2003;:824. (2003)
4. Néveol, A., Rogozan, A., Darmoni, S.J. : Automatic indexing of online health resources for a French quality controlled gateway. In Information Processing & Management, in press. (2005).
5. 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
6. 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.
7. Salton G., and Buckley C., Term weighting approaches in automatic text retrieval. In: Information Processing and Management 24(5) (1988) 513--523.

8. 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.
9. 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.
10. Gaudinat A, Boyer C. Automatic Extraction of MeSH terms from MEDLINEs Abstracts. Workshop on Natural Language Processing in Biomedical Applications, NLPBA2002: 53-57.
11. 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.
12. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004;2004:268-72.
13. Funk ME., Reid CA. and Mc Googan LS. Indexing consistency in MEDLINE. Bull. Med. Libr. Assoc. 71(2):176-183. (1983).
14. Manning, C.D. and Schütze, H. Fondations of Statistical Natural Language Processing (pp. 534-6). MIT Press, Cambridge, MA. (1999)
15. Zweigenbaum P., Baud R., Burgun A., Namer F., Jarousse E., Grabar N., Ruch P., Le Duff F., Thirion B., Darmoni SJ. (2003) UMLF : construction d'un lexique médical francophone unifié. Proceedings of JFIM03, (in press).
16. Névéol A., Rogozan A., Darmoni SJ. Indexation automatique de ressources de santé à l'aide paires de descripteurs MeSH. Submitted to TALN 2005.