Evaluation of French and English MeSH Indexing Systems with a Parallel Corpus

Aurélie Névéol[a, b] , James G. Mork[c], Alan R. Aronson[c], and Stefan J. Darmoni[a, b]

[a]Laboratoire PSI - FRE 2645 CNRS - INSA de Rouen, France
aneveol@insa-rouen.fr
[b]CISMeF - CHU de Rouen, 1, rue de Germont, 76031 Rouen, France
stefan.darmoni@chu-rouen.fr
[c]National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA
{mork,alan}@nlm.nih.gov

Abstract

Objective: This paper presents the evaluation of two MeSH® indexing systems for French and English on a parallel corpus.

Material and methods: We describe two automatic MeSH indexing systems - MTI for English, and MAIF for French. The French version of the evaluation resources has been manually indexed with MeSH keyword/qualifier pairs. This professional indexing is used as our gold standard in the evaluation of both systems on keyword retrieval.

Results: The English system (MTI) obtains significantly better precision and recall (78% precision and 21% recall at rank 1, vs. 37%. precision and 6% recall for MAIF). Moreover, the performance of both systems can be optimised by the breakage function used by the French system (MAIF), which selects an adaptive number of descriptors for each resource indexed.

Conclusion: MTI achieves better performance. However, both systems have features that can benefit each other.

Keywords:

Automatic Indexing, Internet, Medical Subject Headings, Parallel Corpus.

Introduction

The Internet has become a very ubiquitous source of information in numerous fields, including health. Several tools have been developed in order to meet the users' need to find precisely what they are looking for in terms of health information among the numerous documents available online. With PubMed®[1] and the MEDLINE® database, the U.S. National Library of Medicine (NLM) was among the pioneers in medical information retrieval. Today, the MEDLINE database contains 15 million MeSH-indexed resources in English. Since 1995, CISMeF[2] (French acronym of Catalogue and Index of Medical On-Line Resources) has been carrying on similar work on the most important resources of institutional health information in French [1]. It currently contains more than 14,000 resources selected for health professionals (e.g. evidence-based resources - practice guidelines & consensus conferences- and technical reports), medical students (e.g. lecture notes), and patients (e.g. patient education handouts). An average of 55 new resources are added each week. Indexing is a decisive step for the efficiency of information retrieval within both the MEDLINE database and 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 1980s [2]. More recently, MeSH indexing tools have also been available for French. This paper presents the results of a comparative evaluation of two MeSH indexing systems, the Medical Text Indexer (MTI), which is used at the NLM to help with the indexing of English resources [9], and the MesH Automatic Indexer for French (MAIF) which is currently developed within the CISMeF team. The evaluation aims at assessing both the systems performance and the complementarity of the methods implemented.

Materials and Methods

This section introduces the different elements involved in the evaluation, viz. the MeSH indexing systems developed both in the United States and France, the evaluation (parallel) corpus and the evaluation methods.

NLM - Medical Text Indexer (MTI)

MTI results from the combination of two MeSH Indexing methods. These methods are a Natural Language Processing (NLP) approach based on MetaMap Indexing (MMI) and a

[1] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?

[2] http://www.cismef.org

statistical, knowledge-based approach called PubMed Related Citations. MTI combines the results of the two methods by performing a specific post processing task.

(1) The MMI method of discovering Unified Medical Language System® (UMLS®)[3] concepts consists of applying the MetaMap program [3] to a body of text and then scoring and ranking the resulting concepts using a combination of frequency and relevance factors. The UMLS concepts are then refined into MeSH terms using a Restrict to MeSH [4] method of restricting given UMLS concepts to the semantically closest MeSH term.

(2) The PubMed Related Citations method [5] indirectly computes a ranked list of MeSH terms for a given title and abstract by first finding the MEDLINE citations most closely related to the text based on the words they have in common with some adjustment for document length. The final list of MeSH terms are extracted from the MeSH fields of the related citations and are assigned the same score as that of the document.

A clustering algorithm then produces a single ranked list of recommended MeSH terms by combining the recommendations from both methods using term weights, co-occurrence information, and whether the term was found in the title or not. This final ranked list is then subjected to a battery of rules designed to filter out irrelevant indexing recommendations. These filtering rules are based somewhat on NLM indexing policy and also on experience with the data. MTI provides three levels of filtering depending on the balance of precision and recall required for the results. The strict filtering level removes all recommendations that were not supported by both of the MTI methods leaving a small list of very good recommendations with high precision and low recall. The medium filtering level uses the specific strength's of each MTI method to help validate and remove recommendations that are too general or spurious, providing a good sized recommendations list with average precision and recall. The base level of filtering is done regardless of whether strict or medium filtering has been requested and where we apply rules based on actual Indexer rules and developed over time based on familiarity with the data. The base level of filtering provides a reasonable mix of good and bad recommendations with higher recall and lower precision then the other two levels. The rules applied in the basic filtering level focus on five main areas: (1) *addition*, (2) *removal*, and (3) *boosting* of recommended terms based on the other terms in the list, (4) *disambiguating* known problematic recommendations based on a contextual review of the text and other terms in the list, and (5) the *substitution* of subheadings for main headings where applicable.

MeSH Automatic Indexer for French (MAIF)

MAIF is similar to MTI in that it is a combination of two MeSH indexing approaches: an NLP approach, and a statistical, knowledge-based approach. MAIF differs from MTI in the specifics of the two methods:

(1) The NLP approach (detailed in [6]) follows 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 containing full MeSH terms and their variants is used to extract medical concepts. Dictionary entries contain a specific "form" of a MeSH term that is likely to appear in natural language text (i.e., the actual term, its inflected forms, its synonyms or an inflected form of a synonym ...) as well as the MeSH term, itself. Therefore, all the variants of the concepts (inflected forms, synonyms, etc.) are taken into account to compute the frequency of each concept, and each is translated into its corresponding MeSH term. According to MeSH hierarchical information , the occurrences of ancestors are redistributed equally among occurring children in order to increase the score of the most precise terms. As recommended by [7], a tf*idf normalization is then used to compute relevance scores for each MeSH term. Moreover, recurring check tags are promoted to 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 the breakage function described in [6]. The scores assigned to each MeSH candidate represent the likelihood of a candidate to be a good indexing term: the higher the score, the more likely it is that the corresponding MeSH term is good indexing candidate. Given a list of indexing candidates and the score that has been assigned to them, the breakage function is meant to point to indicate a breach of continuity in the scores, therefore highlighting the point in the candidate list where terms become significantly less likely to be correct indexing terms. This point is the "threshold", and the final index for a resource consists of all the terms ranked above this threshold.

(2) The Knowledge Based approach is based on a reference method in the field of classification, namely the *k*-Nearest Neighbour (*k*-NN) method. The underlying principle is very straight-forward. Assuming that a collection of labelled resources C is available, the distance between a new resource r and each resource of C is computed in order to select the k nearest neighbours to r. In our application, the resources are represented by a bag of words constituted by words of the title after stop word filtering. The distance between two resources is represented by the number of common title words.

In the case of a one class classification, the most frequent class among the k neighbours is selected for the new resource. However, the indexing of a resource is composed of a set of MeSH keywords (or keyword/qualifier pairs) which size is unknown. In other words, there is no information on how many keywords (or pairs) should be selected to index a resource. Hence, the indexing of r consists of a set of MeSH candidates to which is assigned a score S (between 1 and k) according to the number of occurrences of the candidate in the indexing of the k neighbours. The final candidate selection is processed with the breakage function described in [6]. The combination of these two approaches in MAIF takes into account the relative score assigned to the terms by each approach. We compute the "relative score" of a term by dividing the score of the term according to the corresponding approach

[3] http://www.nlm.nih.gov/research/umls/about_umls.html (last visited on 03/10/05)

by the sum of all the scores assigned by this approach. Therefore, the score resulting from the two approaches is the sum of the relative scores obtained from each approach. Subsequently, terms are ranked by decreasing scores. However, the terms that have been selected by the two approaches are promoted at the top of the indexing.
Although MAIF 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.

Evaluation corpus and measures

Although the CISMeF catalogue is focusing on referencing French resources, it also includes multilingual resources, if one of the languages the resources are available in is French. Indeed, 1,550 CISMeF resources are available in French and English (about 11%[4]). Most of these resources come from governmental Canadian websites such as Health Canada[5] and the Canadian Pediatric Association[6], which guarantees high quality translation for the documents. The corpus used for this evaluation is composed of 51 resources randomly selected from the Canadian institutional health resources in the CISMeF catalogue. It contains about 270,000 words altogether, which represents about 2 MB. These resources have been manually indexed by five professional indexers in the CISMeF team. In the literature, manual indexing is considered to be the gold standard to which the automatic indexing produced by each system is compared, although the inter-expert variability is high. The average number of isolated keywords used by the indexers to index a resource in the evaluation corpus is 5.86 +/- 5.03. The average number of keywords or keyword/qualifier pairs used to index a resource in the evaluation corpus is 8.78 +/- 7.54.
MTI was originally meant to index Medline citations, which are composed of the title and abstract of an article. Therefore, the text used to produce the automatic indexing is usually less than 300 words long. In the evaluation corpus used for this experiment, the average size of an English resource can be estimated to approximately 2.100 words, which is seven times longer. Therefore, in order to allow for a reasonable processing time, the texts were segmented into sentences to produce about 2.000-character-long chunks. Each chunk was indexed independently with both MTI paths. The results were recombined, and filtering was applied on the recombined results to create a set of indexing recommendations for the resource.
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 [8]. 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, MTI retrieves 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*>.

Results

Table 1 shows the precision and recall (P-R) obtained by each system at fixed ranks 1 through 10. With strict filtering, MTI retrieved more than 10 keywords for only 42 resources, therefore the figures may not be representative of the system performance beyond rank 10. For 10 resources of the evaluation corpus, it was not possible to find 10 neighbors. Therefore, the figures presented in this table concern the 40 resources for which the k-NN method could be applied

Rk	**MTI Strict**	**CISMeF-NLP**
	P – R – F	P – R – F
1	78.43 - 21.24 - 33.42	54.90 - 9.71 - 16.50
3	54.90 - 40.24 - 46.44	40.47 - 28.67 - 33.56
5	40.78 - 45.67 - 43.09	30.59 - 32.63 - 31.57
7	33.08 - 50.78 - 40.06	26.08 - 39.71 - 31.48
10	26.74 - 55.60 - 36.12	22.60 - 49.42 - 31.02
T	**38.84 - 53.28 – 44.90 (T=11.18)**	**38.56 - 35.00 – 36.69 (T=5.24)**
Rk	**MTI Medium**	**MAIF**
	P – R – F	P – R – F
1	74.51 - 20.31 - 31.92	37.25 - 6.14 - 10.54
3	49.65 - 36.00 - 41.74	34.61 - 21.71 - 26.68
5	39.61 - 45.41 - 42.31	26.27 - 26.75 - 26.51
7	31.37 - 49.69 - 38.46	21.82 - 32.14 - 25.99
10	25.49 - 54.59 - 34.75	18.24 - 39.29 - 24.91
T	**36.88 - 59.40– 45.50 (T=15.20)**	**27.20 - 36.06 – 31.01 (T=7.46)**
Rk	**MTI Default**	**CISMeF 10-NN**
	P – R – F	P – R – F
1	74.51 - 20.31 - 31.92	21.95 - 3.49 - 6.02
3	50.29 - 36.20 - 42.10	14.56 - 8.05 - 10.37
5	39.22 - 45.02 - 41.92	13.17 - 12.02 - 12.57
7	31.10 - 48.84 - 38.00	13.95 - 16.61 - 15.16
10	25.10 - 53.22 - 34.11	12.44 - 20.29 - 15.42
T	**33.64 - 61.82 – 43.57 (T=20.90)**	**12.35 - 23.18 – 16.11 (T=11.55)**

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

We have also used the breakage function described in [6]. The last line of Table 1 shows the average precision and recall at the threshold and the average threshold (between brackets).

[4] Figures as of 03/07/05.
[5] http://www.hc-sc.gc.ca (last visited on 03/07/05)
[6] http://www.cps.ca/ (last visited on 03/07/05)

Figure 1 allows a comparison of the two systems through F-measure.

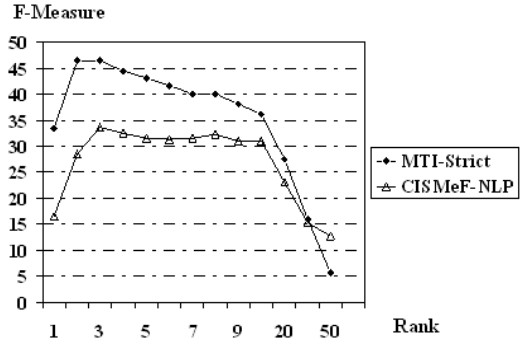

Figure 1: plot of F-Measure vs. fixed ranks for each indexing system.

We can see that the F-measure increases until rank 3 and remains stable until rank 10 for MAIF, while it increases until rank 2 for MTI and slowly decreases until rank 10.

Discussion

We consider the results we obtained to be representative because the evaluation corpus contained more than 30 resources, a minimum for statistical representativity. In addition, both MTI and MAIF have exhibited similar performance in previous evaluations on larger corpora presented in [9] for MTI and [10] for MAIF.

Global performances of the systems

MTI obtains the best overall results both at fixed ranks and at the threshold. As shown in Table 1, MTI achieves a better precision and recall than MAIF/CISMeF at all times. Figure 1 reflects this observation, as MTI's F-measure curve is above that of CISMeF-NLP (except at rank 50).

The difference in performance may result from several factors, related to the experimental condition of MAIF, difference in linguistic resources used, availability of bio-medical resources for French and English, and the method for combining the different approaches.

(1) MeSH coverage by MAIF. The NLP approach used in MAIF currently works with a dictionary covering 60% of MeSH – the comprehensiveness of the dictionary is a key factor in system performance. Used on the same corpus with a previous version of the dictionary covering only 33% of MeSH [10], the NLP approach obtained significantly poorer results. Although it didn't have any influence on this particular study as the evaluation corpus is drawn from the CISMeF collection, it is important to stress that the k-NN approach is also limited in terms of MeSH coverage since the CISMeF catalogue uses about 50% of MeSH.

(2) Difference in Linguistic resources. MTI, on the other hand, is able to cover 100% of MeSH, and also uses UMLS resources for its indexing. The UMLS metathesaurus contains biomedical terms from over 70 terminologies and provides semantic links between alternative names and views of the same concept in order to identify useful relationships between the concepts. Although considerable efforts are made towards increasing the number of linguistic resources available for French in the biomedical domain [10], a significant number of terms (including 50,000 MeSH entry terms) remain to be translated into French.

(3) Combination of NLP and statistical, knowledge based methods. After combining the indexing recommendations obtained from the NLP and statistical approaches, MTI uses a filtering method to enhance the keyword list. As a direct result of this post-processing, precision and recall increase significantly (e.g., at rank 3, precision with default (no) filtering is 50% whereas precision with strict filtering is 54%). MAIF also uses some post-processing on the NLP approach, when indexing rules are applied. However, the impact on the global results is much less significant. Post processing is a decisive step for automatic indexing, as previous work [10] underlined that several MeSH indexing systems tend to retrieve keywords that are either too broad or too narrow to be considered adequate. As some of these mistakes are recurring (e.g. retrieving keywords such as <*disease*> or <*syndrome*>) they may be corrected through post-processing. We can assume that the method used to process lengthy resources with MTI (segmentation of the text, and recombination of the indexing recommendations produced for each chunk) has little influence (besides processing time) on the performance of the system since post-processing is applied on the recombined results, therefore considering the resource as a whole in a fundamental step of the indexing process.

Among the keywords retrieved by MAIF that were not selected by the human indexers, 52% were check tags (at rank 1). The indexing rule to systematically select check tags may be over-enforced in MAIF. However, in a previous study [6] we have shown that there was a significant lack of check tags in CISMeF's manual indexing. Funk et al. [12] have shown that the inter-indexer consistency is at most 70%, which underlines the subjectivity of the indexing task. In fact, a qualitative analysis of the keywords retrieved by both systems but not by the human indexer indicates that these keywords are not irrelevant *per se*. Most of them are either too broad or too narrow to describe the resource adequately. A few keywords are actually relevant, and have been omitted by the human indexer, either by mistake, or due to time constraints in indexing the resource.

The NLP approach used in MAIF performs better alone on this corpus than combined with the kNN approach. This clearly underlines the limits of the kNN approach, whose performance strongly depends on the size of the "knowledge base" used. For MTI, the size of the database is 15 million citations, vs. 14,000 for MAIF. In 10 cases out of 50, it was not possible for MAIF to find 10 neighbors for resources to be indexed. Worse still, in some cases, the "nearest" neighbors in fact deal with different topics, and the resulting indexing is inadequate. Our previous evaluation study [10] shows that the k-NN approach performed better when used for pair indexing. In fact, for pair indexing, the combination of k-NN and NLP

methods in MAIF gave better results than each method used separately.

Added value of the Threshold function

The threshold function is efficient for both MTI and MAIF in terms of maximizing precision. Precision at the threshold is comparatively higher than precision at the equivalent fixed rank (e.g. 39% at threshold 11 vs. 27% at rank 11 for MTI "Strict"). For MAIF, MTI "Default" and "Medium", the F-measure at the Threshold is actually superior to the F-Measure at any given fixed rank (e.g. F-measure at the average threshold 7 is 31 vs. 27 – maximum value at rank 4 for MAIF). For MTI "Strict" however, the F-measure at the threshold is high, but it is not the highest (F-measure at threshold is 44,9 vs. 46,4 – max at rank 2).

Perspective

This comparative study highlights that the two systems have complementary features. MTI performs very well in single keywords retrieval, and also provides a post processing method in order to improve the results further. On the other hand, MAIF is able to retrieve keyword/qualifier pairs almost as efficiently as single keywords, and provides a breakage function to select the optimal number of indexing terms for a given resource. The linguistic resources used by MAIF's NLP approach need to be enriched so as to deal with 100% of MeSH. Moreover, MAIF could integrate MTI's post-processing method to improve its performance. Similarly, we are planning to adapt the pair retrieval technique used by MAIF for English pair retrieval in MTI. The breakage function could be used by both systems for fully automatic indexing.

Conclusion

This paper presents a comparative evaluation of two MeSH indexing systems for French and English through a parallel corpus. MeSH isolated keywords were retrieved by MAIF (French) and MTI (English) from the 50 resources of the evaluation corpus and compared to the manual gold standard. The best precision (78%) is achieved by MTI at rank 1, with Strict filtering. This performance can be explained by a larger MeSH coverage, the use of comprehensive linguistic resources and an efficient keyword filtering method. Future work may involve using MTI's expertise to increase MAIF's performance, and adapting MAIF's specific features (keyword/qualifier pair retrieval and adaptive threshold selection) for MTI.

References

[1] Darmoni SJ, Leroy JP, Thirion B, Baudic F, Douyère M and Piot J. CISMeF: a structured Health resource guide. Meth Inf Med 2000: 39(1): 30-5.

[2] Humphrey SM, and Miller NE. Knowledge-based indexing of the medical literature: The Indexing Aid Project. J Am Soc Inf Sci 1987 May: 38(3): 184-96.

[3] Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp.* 2001;:17-21.

[4] Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *Proc AMIA Symp.* 1998:815-9.

[5] Kim W, Aronson AR, Wilbur WJ. Automatic MeSH term assignment and quality assessment. Proc AMIA Symp. 2001;:319-23.

[6] Névéol 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).

[7] Salton G., and Buckley C., Term weighting approaches in automatic text retrieval. In: Information Processing and Management 24(5) (1988) 513--523.

[8] Manning, C.D. and Schütze, H. Fondations of Statistical Natural Language Processing (pp. 534-6). MIT Press, Cambridge, MA. (1999)

[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] Néveol A., Mary V., Gaudinat A., Rogozan A., Boyer C., Darmoni S.J. : A Benchmark evaluation of the French MeSH indexing systems. Proc. AIME 2005 (in press).

[11] 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.

[12] Funk ME., Reid CA. and Mc Googan LS. Indexing consistency in MEDLINE. Bull. Med. Libr. Assoc. 71(2):176-183. (1983).

Address for correspondence

Aurélie Névéol
Equipe CISMeF, CHU de Rouen
1, rue de Germont – 76031 Rouen, FRANCE
E-mail: aneveol@insa-rouen.fr