Multi-terminology indexing for the assignment of MeSH descriptors to medical abstracts in French

Suzanne Pereira, PhD [1], Saoussen Sakji, MSc [2], Aurélie Névéol, PhD [3], Ivan Kergourlay [2], Gaétan Kerdelhué [2], Elisabeth Serrot, PhD [1], Michel Joubert, PhD [4], Stéfan J. Darmoni, MD, PhD [2]

[1]VIDAL, Issy les Moulineaux, France [2]CISMeF, LITIS EA 4108, University of Rouen, France [3]NLM, Bethesda, USA and [4]LERTIM, Marseille Medical University, France

Abstract

***Background**: To facilitate information retrieval in the biomedical domain, a system for the automatic assignment of Medical Subject Headings to documents curated by an online quality-controlled health gateway was implemented. The French Multi-Terminology Indexer (F-MTI) implements a multi-terminology approach using nine main medical terminologies in French and the mappings between them. **Objective**: This paper presents recent efforts to assess the added value of (a) integrating four new terminologies (Orphanet, ATC, drug names, MeSH supplementary concepts) into F-MTI's knowledge sources and (b) performing the automatic indexing on the titles and abstracts (*vs. *title only) of the online health resources. **Methods**: F-MTI was evaluated on a CISMeF corpus comprising 18,161 manually indexed resources. **Results**: The performance of F-MTI including nine health terminologies on CISMeF resources with Title only was 27.9% precision and 19.7% recall, while the performance on CISMeF resources with Title and Abstract is 14.9 % precision (-13.0%) and 25.9% recall (+6.2%). **Conclusion:** In a few weeks, CISMeF will launch the indexing of resources based on title and abstract, using nine terminologies.*

Introduction

In the biomedical domain, CISMeF (French acronym for Catalogue and Index of Online Health Resources in French) helps health professionals, patients and medical students to find quality electronic information available on the Internet [1]. The catalogue describes prominent quality health resources (N=65,439) with 11 out of 15 metadata from the Dublin Core (DC) set [2] including title, abstract and resource types. In addition, resources are indexed with a set of indexing terms describing the information content. The indexing terms consist of descriptor/qualifier pairs or descriptors from the MeSH® thesaurus (Medical Subject Headings), the U.S. National Library of Medicine's (NLM's) controlled vocabulary used to index articles from the MEDLINE® bibliographic database.

Faced with the growing amount of online resources to be indexed and included in the catalogue, the CISMeF team consistently evaluated advanced automatic MeSH indexing techniques [3]. In August 2006 this project led to the effective use of a bag-of-words algorithm to automatically index "low priority" resources to be included in CISMeF.

After the previous F-MTI study in 2008 [4], the automated indexing tool to index health resources in CISMeF was upgraded from a bag-of words algorithm based on a mono-terminology approach to F-MTI, which uses several health terminologies. In 2008, besides MeSH, four health terminologies were included in F-MTI: ICD-10 (International Classification of Diseases) and SNOMED 3.5 (Systematized Nomenclature of medicine) which are included in the UMLS® metathesaurus, CCAM (the French equivalent of US CPT) and TUV (a French terminology for therapeutic and clinical notions for the use of drugs) which are not included in the UMLS (yet). These four terminologies are mapped to the French version of MeSH; mappings from the latest two (CCAM and TUV) have been produced manually as part of the effort to implement F-MTI.

Since then, F-MTI has enabled the fully automatic indexing of 33,986 resources and the semi-automatic indexing (i.e. automatic indexing followed by revision by a human indexer) of another 12,440 resources based on resources titles. The most important resources (in particular clinical guidelines) are still indexed manually, without any use of automatic indexing tools (N=19,013).

Work on the 2008 European Union funded PSIP project devoted to optimizing patient safety during drug prescription led to the addition of four new terminologies devoted to drugs into F-MTI's knowledge sources: the Anatomical Therapeutic Chemical (ATC) Classification (N=5,514), drug names with international non-proprietary names (INN) and brand names (N=11,353), the Orphanet

thesaurus for rare diseases (N=7,424) and the chemical substances and pharmacological action terms of the MeSH Supplementary Concepts translated into French by the CISMeF team (N=6,505 out of over 180,672).

One of the remaining challenges that the CISMeF automatic indexing algorithm needs to address is to evaluate the added value of indexing based on the title and abstract of resources (*vs.* title only), which has proved successful with the US National Library of Medicine (NLM)'s Medical Text Indexer (MTI). MTI processes the title and abstract (if available) of articles to be indexed for MEDLINE [5][1].

Because the number of terminologies in F-MTI has grown and will continue to grow, the CISMeF team has drastically improved F-MTI in terms of response time, including using a multi-threaded version. As a result, it is now possible to perform indexing based on resource title and abstract (vs. title only) using F-MTI.

The goal of this study is to evaluate the potential added value of integrating new terminologies in F-MTI and expanding the automatic indexing from resource titles to resources titles and abstracts.

Materials and Methods

The MeSH automatic indexing tool used in this study, F-MTI, implements a bag-of-words indexing algorithm previously described [9]. In this section, we describe how the original indexing algorithm was improved by the addition of four terminologies to the workflow.

Details on the four terminologies: (1) The Anatomical Therapeutic Chemical (ATC) Classification, controlled by the Collaborating Centre for Drug Statistics Methodology of the World Health Organization, is used to classify drugs (N=5,514). Drugs are divided into various groups according to the organ or the system on which they act and/or their therapeutic and chemical characteristics. In ATC, drugs are classified in five groups at different levels. Each level of the classification corresponds to an ATC code and an ATC label. The label of the 5th level corresponds to the International Generic Name of the substance, when it exists (e.g. C03CA01 Furosemide).

(2) Orphanet provides rare disease information to healthcare professionals, patients, and their relatives, with the goal of improving diagnosis, care and treatment. Orphanet has developed a thesaurus for rare diseases (e.g. adult ataxia), available in five European languages (English, French, Spanish, German & Italian) (N=7,424).

(3) Within the MeSH thesaurus, the MeSH Supplementary Concepts (SC) are terms of reference, describing mostly chemical substances. Unlike main headings and subheadings they are not organized hierarchically. However, SC are semantically related to MeSH descriptors. For each SC, a mapping to one or several MeSH descriptors is available (e.g. the SC lombazole is mapped to the descriptor imidazoles). In addition, links to pharmacologic actions headings are also available, when relevant. The CISMeF team has translated the SCs corresponding to the main drugs and toxic chemical substances for Humans into French (N=6,505 SCs out of over 180,672).

(4) The Brand names and International Drug Names classifications were provided by the Vidal company (N=11,353). The brand name of a drug corresponds to the name of the drug specifically attributed by the manufacturer (e.g. Niquitin®; N=8,711). International Drug Names (or International Non-Proprietary Names –INN-) is a classification of drugs according to their active ingredients (e.g. nicotin; N=2,642).

F-MTI multi-terminology process: For each resource, the title/subtitle and/or abstract is first broken into sentences. Then each sentence is normalized (accents are removed, all words are switched to lower case and stemmed …) and stop words are removed to form a bag of words containing all the content words. The "bag" thus obtained is matched independently of the order of the words against all the MeSH, ICD10, SNOMED Int, CCAM, TUV, ATC, INN, Brand names, Orphanet terms that have been processed in the same way. All terms containing at least one word of the sentence are retrieved. Longer matches are preferred to shorter ones. All these candidate indexing terms are restricted to the semantically closest MeSH terms using inter-concept relationships. Some mappings are taken from the UMLS[2] (ICD10-MeSH and SNOMED-MeSH). They include mappings between MeSH supplementary concepts and one or more MeSH descriptors. Other mappings were manually created by several experts from Orphanet, Vidal and CISMeF (Orphanet-MeSH, CCAM-MeSH, ATC-MeSH and TUV-MeSH).

The resulting pool of MeSH indexing terms includes terms that were obtained directly and terms obtained indirectly using the relationships. Finally, when both

[1] F-MTI was named after the US NLM MTI.

[2] Specifically, the 2007AB release

MeSH descriptors and qualifiers are retrieved, all the legal descriptor/qualifier pairs are formed.

Table 1 shows a sample resource from the corpus, along with manual indexing and automatic multi-terminology indexing produced by F-MTI.

Title & ***Abstract***	Thesaurus of Public health version 3.02 Topic access, internal search engine for the topic access.
Manual (Gold Standard) MeSH indexing	* vocabulary, controlled * public health * terminology as topic
Multi-terminology indexing	***MSH*** public health ***MSH*** vocabulary, controlled ***MSH*** research **SNMI** version ***SNMI*** internal ***SNMI*** internist' ***SNMI*** one

Table 1. Manual and automatic indexing of a CISMeF resource. In the last row, the terminologies are shown in bold.

Execution time: F-MTI is implemented in Perl. To reduce F-MTI's run-time, one research engineer (IK) optimized the F-MTI code rewriting of some portions of the algorithm. Improvements included adjustments to the loading process of the various word tables and dictionaries used in the bag of words algorithm (e.g. list of stop words, list of non pertinent words). These steps have led to a reduction of time spent by a factor 3 to 10 (depending on the length and type of sentences). Then, F-MTI was ported to a much more powerful machine with enough memory to avoid to swap (a multi-core server (n=8) and 8 Go of RAM). The algorithm was parallelized, so that up to 8 resources may be process simultaneously. Overall, the processing time for all corpus titles was reduced by a factor 90. Finally, it was possible to run F-MTI on the entire CISMeF manually-indexed corpus processing both titles and abstracts.

Test Corpus: F-MTI was evaluated on a CISMeF corpus of 18,161 manually indexed resources comprising at least a title, a subtitle and an abstract. For each resource in the corpus, the indexers selected the title, the subtitle and wrote a short abstract. They also described the resource by selecting the resource types and a set of MeSH indexing terms. In other words, indexers selected descriptors and descriptor/qualifier pairs from the 24,765 descriptors and 83 qualifiers available in the 2008 MeSH thesaurus and assigned to each a "major" or "minor" weight depending on how substantively the concept represented by the indexing term was discussed in the resource.

Evaluation measures: Precision, recall and F-measure were calculated to show the performance of F-MTI indexing compared to the reference manual indexing. Three variants of F-MTI indexing were assessed: (a) Indexing on titles and subtitles using 5 terminologies (results from [4]) (TST/FMTI5), (b) Indexing on titles and subtitles using 9 terminologies (TST/FMTI9) and (c) Indexing on titles, subtitles and abstracts using 9 terminologies (TSTA/FMTI9).

Precision is the number of indexing terms present in both the candidate and gold standard sets divided by the total number of indexing terms in the candidate set. It measures how well the bag-of-words algorithm weeds out what is not identical to the gold standard (manual indexing). *Recall* is the number of indexing terms present in both the candidate and gold standard sets divided by the total number of indexing terms in the gold standard set. It measures how well gold standard indexing terms were extracted. *F-measure* is the harmonic mean of precision and recall. *F-measure* =(2**Precision***Recall*)/(*Precision*+*Recall*)

In addition, we considered the performance obtained on three categories of terms:

- Indexing Terms: MeSH descriptors or descriptor/qualifier pairs (e.g "Asthma", "Breast Neoplasms/prevention and control").
- Descriptors: MeSH descriptor - qualifiers are not taken into account (e.g. in the pair "Breast Neoplasms/prevention and control" only the descriptor "Breast Neoplasms" will be considered).
- Central-concept Descriptors: Only major MeSH descriptors labeled with the star symbol "*" without qualifiers are taken into account (e.g. *Pharyngitis).

Results

Table 2 displays F-MTI results in three different contexts: (a) the results with five terminologies on titles only - presented at AMIA 2008; (b) the results with nine terminologies on titles only; and (c) the results with nine terminologies on titles and abstracts.

As shown in Table 2 ("Descriptors" line, column a vs. b), F-MTI exhibits higher precision when using only five terminologies: 35.5% vs. 34.8% when nine terminologies are used. On the other hand, a higher recall is obtained when using nine terminologies: 28.2% vs. 23.1% when using only five terminologies.

	Performance Precision (%) – Recall (%) (F-measure) (%)		
	(a)TST/FMTI5	(b)TST/FMTI9	(c)TSTA/FMTI9
Indexing terms	25.9 – 13.5 (17.8)	27.9 – 19.7 (23.1)	14.9 – 25.9 (18.9)
Descriptors	35.5 – 23.1 (28.0)	34.8 – 28.2 (31.2)	18.3 – 37.5 (24.6)
Central-concept Descriptors	30.5 – 38.1 (33.9)	29.3 – 44.2 (35.2)	12.7 – 51.5 (20.4)

Table 2. Performance of F-MTI compared to the human indexers on a test corpus.

A similar trade-off is observed (Table 2, "Descriptors" line, column b vs. c) when comparing indexing on title only vs. title and abstract. A better precision is obtained with indexing on title only: 34.8% (vs. 18.3% for title and abstract) whereas a better recall is obtained with indexing on title and abstract: 37.5% (vs. 28.2% for title only).

Overall, the F-measure is higher for Titles and subtitles indexing using nine terminologies (23.1%, 31.2% and 35.2% compared to 17.8%, 28.0% and 33.9% using five terminologies and 18.9%, 24.6% and 20.4% for titles, subtitles and abstract indexing).

Discussion

Added value of integrating new terminologies: The performance of F-MTI including five health terminologies on CISMeF manually-indexed resources (n=18,161) with Title only was 25.9% precision and 13.5% recall, while the performance with nine terminologies was 27.9% precision (-2%) and 19.7% recall (+6.2%). The use of nine terminologies instead of five terminologies allows exploiting a bigger semantic network. Access to a bigger semantic network implies that more concepts may be extracted because of more terms, variants and synonyms important for natural language processing.

Added value of indexing abstracts: On indexing terms, the performance of F-MTI including nine health terminologies on CISMeF manually-indexed resources (n=18.161) with Title only was 27.9% precision and 19.7% recall, while the performance on CISMeF resources with Title and Abstract was 14.9 % precision (-13.0%) and 25.9% recall (+6.2%). The lower precision is partly due to the high number of MeSH terms extracted (an average of 10(+/-7) terms by F-MTI and 6(+/-7) for manual indexers) when the abstract is considered (vs. an average of 3(+/-2) terms by F-MTI and 6(+/-7) for manual indexers without the abstract). Future work should address the development of a filtering method aiming at reducing the number of erroneous terms in the final set.

These figures were satisfactory for the four CISMeF indexers because in their daily practice on supervision, they are expecting the highest recall, knowing that the precision will diminish. These results can be compared to those of MTI Overall Statistics for the 2007 Indexing Year [12], which show that the performance on citations with Title Only was 28.08% precision and 29.91% recall, while the performance on citations with Title and Abstract was 30.23% precision (+2.20%) vs. 53.79% recall (+13.88%). The impact on indexing abstract is then greater with MTI (improvement of both precision and recall) than with F-MTI (improvement of the recall). However, note that the MTI data reported above was obtained on two distinct sets of citations, one set where both title and abstracts were available and one set where only the title was available. This is a major methodological difference with the study presented here where F-MTI was applied to the title and abstract of resources and then to the title only of the same resource set. Note that that the CISMeF abstracts are shorter than MEDLINE abstracts (on average, 38 words vs. 178).

Perspectives: Several other health terminologies available in French will be added to F-MTI in 2009: ICPC (International Classification of Primary Care), DRC (Dictionary of Consultation Results), ICF (International Classification of Functioning, Disability and Health) and WHO-ART (The Who Adverse Reaction Terminology).

Moreover, the CISMeF backoffice in cooperation with a Health Multi-Terminology Server [13] has recently integrated all the terminologies available in F-MTI. In a few weeks, the Doc'CISMeF search engine will enable multi-terminology indexing as well as searching.

In a future study, we will evaluate the added value of full text indexing as MTI has been evaluated to for MEDLINE articles indexing [14].

Conclusion

We developed a multi-terminology automatic indexing tool, the French Multi-Terminology Indexer (F-MTI). This tool is going to be integrated in an online quality-controlled health gateway, CISMeF, to index the resources by title and abstract using several terminologies.

Acknowledgments

This research was supported in part by VIDAL (http://www.vidal.fr/) and the PSIP (Patient Safety through Intelligent Procedures in medication) project from the 7th Framework Program of the European Union ICT-1-5.2 Risk Assessment and Patient Safety program (Grant agreement n° 216130). This work was also supported in part by the Intramural Research Program of NIH, National Library of Medicine. The authors would like to thank CISMeF indexers for their help in the study design and result analysis.

References

1. Douyère M, Soualmia LF, Névéol A, Rogozan A, Dahamna B, Leroy JP, Thirion B, Darmoni SJ. Enhancing the MeSH thesaurus to retrieve French online health resources in a quality controlled gateway. Health Info Libr J 2004;21(4):253-61.
2. Dekkers M, Weibel S. State of the Dublin Core Metadata Initiative. D-Lib Magazine 2003;9(40).
3. Névéol A; Rogozan A, Darmoni S. Automatic indexing of online health resources for a French quality controlled gateway. Information Processing & Management 2006;42(3):695-709.
4. Pereira S, Névéol A, Kerdelhué G, Serrot E, Joubert M, Darmoni SJ. Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue. AMIA Annu Symp Proc. 2008 Nov 6:586-90.
5. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004;2004:268-72.
6. Dutoit D, Nugues P, De Torcy P. The Integral Dictionary: A Lexical Network Based on Componential Semantics. Lecture Notes in Computer Science 2003;2667:368-377.
7. Jacquemin C. Flemm: Un analyseur Flexionnel de Français à base de règles. Traitement automatique des Langues pour la recherche d'information. (éds). Paris: Hermes 2000 :523-47.
8. Bodenreider O, Nelson SJ, Hole WT and Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp 1998:815-9.
9. Névéol A, Pereira S, Kerdelhué G, Dahamna B, Joubert M, Darmoni SJ. Evaluation of a Simple Method for the Automatic Assignment of MeSH Descriptors to Health Resources in a French Online Catalogue. Stud Health Technol Inform 2007:129:407-411.
10. Pereira S, Massari P, Buemi A, Dahamna B, Serrot E, Joubert M, Darmoni SJ. Evaluation of two French SNOMED indexing systems with a parallel corpus. KR-Med 2008 [Poster in Press].
11. Pereira S, Massari P, Joubert M, Serrot E and Darmoni SJ. Exploring Multi-terminology Indexing of Discharge Summaries. Stud Health Technol Inform 2008 [Poster in Press].
12. Aronson, A.R, Mork, J.G., Lang, F.M, Rogers, W.J., & Névéol, A. (2008). NLM Medical Text Indexer: a tool for automatic and assisted indexing. NLM Technical Report No. LHNCBC-TR-2008-002. Bethesda, MD: U.S. National Library of Medicine, April 2008.
13. Joubert M, Dahamna B, Delahousse J, Fieschi M, Darmoni SJ. SMTS®: Un Serveur Multi-Terminologies de Santé. Journées Francophones d'Informatique Médicale, 2009 (in press).
14. Gay SW, Kayaalp Mehmet, Aronson AR. Semi-Automatic Indexing of Full Text Biomedical Articles. AMIA Annu Symp Proc. 2005; 2005: 271–275.

Address for correspondence

Suzanne Pereira
VIDAL
21 rue Camille Desmoulins
92420 Issy les Moulineaux
Email: Suzanne.pereira@vidal.fr