Using multi-terminology indexing for the assignment of MeSH descriptors to health resources in a French online catalogue

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

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

Abstract

***Background**: To assist with the development of a French online quality-controlled health gateway (CISMeF), an automatic indexing tool assigning MeSH descriptors to medical text in French was created. The French Multi-Terminology Indexer (F-MTI) relies on a multi-terminology approach involving four prominent medical terminologies and the mappings between them. **Objective**: In this paper, we compare lemmatization and stemming as methods to process French medical text for indexing. We also evaluate the multi-terminology approach implemented in F-MTI. **Methods**: The indexing strategies were assessed on a corpus of 18,814 resources indexed manually. **Results**: There is little difference in the indexing performance when lemmatization or stemming is used. However, the multi-terminology approach outperforms indexing relying on a single terminology in terms of recall. **Conclusion**: F-MTI will soon be used in the CISMeF production environment and in a Health MultiTerminology Server in French.*

Introduction

CISMeF (French acronym for Catalogue and Index of Online Health Resources in French) describes and indexes prominent quality health resources in French to help health professionals, patients and students find medical information online [1]. In the catalogue, resources are described using Dublin Core (DC) metadata [2] including "title" and "resource types". A set of indexing terms is also used to describe the topics discussed in resources. Resources types (RT) are a generalization of the publication types used in MEDLINE®. The indexing terms are descriptors and descriptor/qualifier pairs from the MeSH® thesaurus (Medical Subject Headings), a controlled vocabulary developed by the U.S. National Library of Medicine (NLM) to index articles from the biomedical literature. As defined by the DC Metadata Initiative RTs are used to describe the nature of a resource, whereas MeSH terms describe the subject matter of a resource. As MeSH descriptors and qualifiers, CISMeF RTs are organized in a hierarchical structure. The RTs hierarchy was built manually and has been maintained by the CISMeF team since 1997.

Faced with the growing amount of online resources to be indexed and included in the catalogue, the CISMeF team has been evaluating advanced automatic MeSH indexing techniques [3-4]. 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. Low priority resources include teaching resources or resources discussing topics substantively covered in the catalogue that do not require in-depth indexing. Since then, the bag-of-words algorithm has enabled the automatic indexing of 16,725 resources and the semi-automatic indexing (i.e. automatic indexing revised by a human indexer) of another 6,644 resources based on resource titles. Resources that were indexed automatically are displayed in the catalogue after those that were indexed manually.

After reaching this milestone in automatic indexing, CISMeF has strived to improve the automatic indexing algorithm and make it on par with manual indexing. One of the challenges that need to be addressed is the identification of all the different forms that a term can take in natural language, specifically with respect to lexical and grammatical variations. Most terminologies such as MeSH provide synonyms and variants of terms but this information is usually insufficient to describe all the forms that can be encountered for a given term in a document. For this reason, a number of stemming (carry algorithm, French stemmer developed in the context of the Lucene project[1]) and lemmatization (Sémiographe by Memodata [5], Flemm [6]) algorithms attempt to reduce a word to its base form, respectively its stem or its lemma. This ensures that all lexical forms of a particular term of a terminology can be located within a sentence. Lemmatization is closely related to stemming. In a lemmatized sentence, words are represented by their infinitive form for verbs or their nominative singular form for

[1] http://lucene.apache.org/

nouns. In a stemmed sentence words are stripped of their suffixes.

Lemmatization and stemming are useful to identify variants of lexical forms but they cannot deal with synonyms. To address this issue, researchers have exploited the information available in other medical terminologies besides MeSH. For instance, the NLM's Medical Text Indexer (MTI) [7] used to provide indexing recommendations in English for articles to be included in MEDLINE partly relies on the Unified Medical Language System® (UMLS®) to extract MeSH descriptors. The UMLS Metathesaurus contains 3.6 million different term forms in English (*vs.* 126,000 term forms in French) from over 100 source vocabularies (including MeSH). These concepts are linked by 10 millions relations. In MTI's Natural Language Processing path, UMLS candidate terms are extracted by MetaMap [8] and restricted to the semantically closest MeSH terms using synonymy, interconcept relationships and categorization [9].

A similar approach was implemented in the French Multi-Terminology Indexer (F-MTI). We used four prominent medical terminologies in French mapped to the French version of MeSH to increase the recall of our bag-of-words algorithm: ICD-10 (International Classification of Diseases) and SNOMED 3.5 (Systematized Nomenclature of medicine) which are included in the UMLS, CCAM (the French equivalent of U.S. CPT) and TUV (a French terminology for therapeutic and clinical notions related to the use of drugs).

Our goal in the experiments reported below is two-fold: in the specific context of French MeSH indexing in CISMeF, we want to determine which of lemmatization or stemming performs best with the bag-of-words indexing algorithm. Second, we also aim at evaluating a multi-terminology version of the bag-of-words algorithm.

Materials and Methods

Bag-of-words indexing algorithm: MeSH automatic indexing in CISMeF is performed by a bag-of-words indexing algorithm. This algorithm is applied only to the title of the resources to extract the major indexing terms. In a previous study, bag-of-words indexing was found to extract 58% of the major concepts from the title of 99 teaching resources [10]. The title, URL, editor and date of the resource is manually entered by indexers before the bag-of-words algorithm is applied as described in the next paragraph.

F-MTI mono-terminology process: After the title has been normalized (accents are removed, all words are switched to lower case…) and stop words have been removed, a bag of words containing all the content words is formed. Each word is stemmed or lemmatized. The "bag" thus obtained is matched independently of the order of the words against all the MeSH terms that have been processed in the same way. All terms matching at least one word in the bag are retrieved. Longer matches are preferred to shorter ones. Finally, when both MeSH descriptors and qualifiers have been retrieved, all the legal descriptor/qualifier pairs are formed.

Lemmatization and stemming: Two methods of word normalization have been added to the base processing described above.

After all the words in a title have been normalized a stemming algorithm developed by our team is applied to each word using an ordered table of 63 suffixes to be removed. For example, processing the title "Echographie obstétricale" ("prenatal ultrasonography" in English) results in the following bag of words: "echograph; obstetric".

Alternatively, lemmatization is performed using the Sémiographe. The Sémiographe consists of a dictionary and a semantic network in French [5]. The words are lexically labeled before being assigned their lemmas. For example, processing the title "Echographie obstétricale" results in the following bag of words: "échographie; obstétrical".

F-MTI multi-terminology process: After the title has been normalized, stop words removed, and each word stemmed or lemmatized, the "bag" obtained is matched independently of the order of the words against all the MeSH, ICD10, SNOMED, CCAM and TUV terms that have been processed in the same way. The indexing candidate terms obtained are restricted to the semantically closest MeSH term(s) using interconcept relationships (ICD10-MeSH and SNOMED-MeSH mappings). As a result, the final list of indexing terms consist of MeSH terms obtained directly and MeSH terms obtained indirectly using the interconcept relationships.

Test Corpus: The algorithm was evaluated on a CISMeF corpus comprising 18,814 resources indexed manually by four professional indexers. For each resource in the corpus, indexers selected the title, the resource types and a set of MeSH indexing terms. In other words, indexers selected descriptors and descriptor/qualifier pairs from the 24,357 descriptors and 83 qualifiers available in the 2007 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: The performance of F-MTI was assessed using precision and recall based on the gold-standard indexing provided by CISMeF indexers. Four variants of F-MTI were assessed:

(a) mono-terminology and stemming algorithm
(b) mono-terminology and lemmatization algorithm
(c) multi-terminology and stemming algorithm
(d) multi-terminology and lemmatization algorithm.

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 the ratio of signal to noise. *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.

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

- Indexing Terms (IT): MeSH descriptors or descriptor/qualifier pairs (*e.g* “asthma“, “breast tumors/prevention and control“).
- Descriptors (D): MeSH descriptors, regardless of the qualifiers attached to them (*e.g.* in the pair “breast tumors/prevention and control” only the descriptor “breast tumors“ is considered). For descriptors, we evaluated the indexing performance on three different resource types which represent the three target audiences of the CISMeF catalogue (health professionals, students, patients). These three resource types are, respectively: “guidelines“, “teaching resources” and “patient information”. We have chosen to take resource types into account for descriptors only because most other MeSH indexing tools do not extract descriptor/qualifier pairs.
- Central-concept Descriptors (*D): Only major MeSH descriptors labeled with the star symbol “*” without qualifiers are taken into account (e.g. *Pharyngitis).

To adequately assess the added value of the multi-terminology bag-of-words algorithm, all the indexing terms extracted by F-MTI for 1,000 resources from the manually indexed corpus were analysed by a CISMeF indexer who was asked to rank each indexing term according to its potential impact in the context of information retrieval: "positive impact", "negative impact" or "minor impact". The ranked list of indexing terms was obtained from the multi-terminology algorithm after removing MeSH indexing terms that were already in the gold standard. The 1,000 resources were randomly selected taking into account the respective proportion of guidelines, teaching and patient resources in the corpus.

Results

Comparison between lemmatization and stemming: For indexing terms comparison, the results for the stemming algorithm show a higher precision (29.4% vs. 28.3%) and a higher recall (13.0% vs. 12.1%) compared to the lemmatization algorithm for F-MTI mono-terminology (see table 1). For the F-MTI multi-terminology, the results for the stemming algorithm show a lower precision (25.9% vs. 26.7%) and a higher recall of (13.5% vs. 13.1%) compared to the lemmatization algorithm (see table 2).

	Performance Precision (%) – Recall (%)		
		(a) Mono/stem	(b) Mono/lemma
IT	All	29.4 - 13.0	28.3 - 12.1
D	All	37.7 - 21.3	38.8 - 20.7
	Guidelines	43.7 - 17.9	47.4 - 16.9
	Teaching	51.6 - 24.7	51.9 - 24.8
	Patient	42.4 - 27.5	43.7 - 25.9
*D	All	36 – 36.4	37.7 - 35.6

Table 1. Performance of bag-of-words indexing using mono-terminology (CISMeF corpus with distinction between teaching corpus, guidelines corpus and patient corpus).

Performance of the bag-of-words indexing using multi-terminology: Comparing F-MTI multi-terminology and mono-terminology with the stemming approach, the results show a lower precision (25.9% vs. 29.4%) and a higher recall (13.5% vs. 13.0%) (see table 1 and 2). For the lemmatization approach, the results show a lower precision (26.7% vs. 28.30%) and a higher recall (13.1% vs. 12.1%).

When taking into account the resource types (teaching, guidelines and patient), variations are important: 44.4% in precision and 25.7% in recall for teaching resources, 39.9% in precision and 18.7% in recall for guidelines, and 38.3% in precision and

27.8% in recall for patient resources for the multi-terminology and stemming algorithm. These variations should be related to the average number of MeSH terms assigned manually for each resource type: 5.5 for teaching resources (vs. F-MTI : 2.1), 9.3 for guidelines (vs. F-MTI : 2.9), 3.5 for patient information (vs. F-MTI : 1.5).

Comparing the results for indexing terms, descriptors and major descriptors, we found that F-MTI better extracts major descriptors than indexing terms and finally descriptors. For the major descriptor extraction using stemming and multi-terminology, the precision is 30.5% and the recall is 38.1% for the multi-terminology and stemming algorithm.

		Performance Precision (%) – Recall (%)	
		(c) Multi/stem	(d) Multi/lemma
IT	All	25.9 - 13.5	26.7 - 13.1
D	All	35.5 - 23.1	26.8 - 22.4
	Guidelines	39.9 - 18.7	42.3 - 17.3
	Teaching	44.4 - 25.7	45.7 - 24.4
	Patient	38.3 - 27.8	38.9 - 26.4
*D	All	30.5 - 38.1	31.5 - 37.6

Table 2. Performance of bag-of-words indexing using multi-terminology (CISMeF corpus with distinction between teaching corpus, guidelines corpus and patient corpus).

Added value of F-MTI: The analysis of the F-MTI automatic indexing for 1,000 resources by a CISMeF indexer showed that 4.5% of the descriptors automatically assigned that were not in the manual set in our study were considered as having a positive impact, 79.6% a negative impact and 15.9% a minor impact.

Discussion

Comparison between lemmatization and stemming: The results show that the performance of lemmatization and stemming is very close for both precision and recall. However, lemmatization gives a better precision but a lower recall because of the under-analysis of variant forms.

Performance of the bag-of-words indexing using multi-terminology: The performance of F-MTI using mono-terminology *vs.* multi-terminology is also close in precision and recall on the descriptors.

The use of multi-terminology indexing instead of mono-terminology indexing allows exploiting the semantic network of several terminologies instead of a single one. Access to a bigger semantic network implies that more concepts may be extracted. The results show a better recall for the multi-terminology algorithm but a lower precision compared to mono-terminology. The lower precision is due to mapping errors that are independent of F-MTI. It is important in this study that the mappings link only concepts that have exactly the same meaning. MeSH-ICD-10 and SNOMED-ICD10 mappings from the UMLS were reviewed, as well as mappings produced by SFINM. The reviews evidenced a significant number of meaning and granularity differences between linked concepts. After removing the mappings errors, we hope to obtain a better precision.

Impact on CISMeF indexing procedure: CISMeF's policy is to provide users with a few targeted quality resources rather than a large amount of resources requiring further weeding out by the user. With respect to indexing, this translates in favoring precision over recall. Therefore, based on the results of this study, lemmatization should be used in F-MTI. However, technical considerations cannot be overlooked. Lemmatization requires twice the execution time needed by stemming. In addition, the Sémiographe greatly increases the complexity of the indexing system. In practice, the gain in precision obtained with lemmatization is not significant enough to warrant the cost of the increased technical complexity of the algorithm. Therefore, the CISMeF team has decided to keep stemming as the reference method to be used in a production environment.

The retrospective analysis performed by a CISMeF indexer highlights the relative interest of F-MTI to help the indexers improve manual indexing. 4.5% of the MeSH descriptors evaluated were considered as having a positive impact on information retrieval. These terms were not assigned manually by the indexers but should have been. In this way, the system can help manual indexers to improve their indexing. 15.9% of the descriptors evaluated were considered as having a minor positive impact and could have been assigned to the resources in addition to the manual indexing. We found that 79.6% of the terms extracted by F-MTI considered as noise (because not indexed by the human indexers) were effectively noise and had a negative effect. So we can assume that the precision of F-MTI's indexing is in fact better that what we had measured. We have planned several changes to improve F-MTI's performance, including correcting conceptual mappings, using contextualization and indexing rules… Methods to rank terms will help minimize indexing noise. Moreover the multi-terminology approach will be useful in the future to index the resources with terms from terminologies other than MeSH.

F-MTI including multi-terminology and stemming algorithm will be implemented in the CISMeF production environment in the near future.

Perspectives: To our knowledge, F-MTI is the first multi-terminology tool available for a language other than English. Unfortunately, there are far less medical terminologies available in French (10) than in English (100). As a result, there are fewer UMLS semantic network mappings between MeSH and other terminologies in French.

Through a continuing collaboration with NLM, we are planning to implement the MetaMap algorithm in French to improve on F-MTI's performance.

The work presented in this paper partakes from a larger-scale project called Health MultiTerminology Server in French. The medical terminologies available in French (mostly translated from English) are included in this server, which will be interfaced with F-MTI to interpret users' information queries in the various terminologies. F-MTI will also be used to index CISMeF resources with multiple terminologies, following successful trials applying F-MTI to the coding of patient discharge summaries in electronic patient record with SNOMED [11] and ICD10 [12].

F-MTI will soon be evaluated to index French Summaries of Products Characteristics using the Unified Vidal Thesaurus.

Conclusion

We developed a MeSH automatic indexing tool, the French Multi-Terminology Indexer (F-MTI). F-MTI with the multi-terminology and stemming algorithm will soon be implemented in the CISMeF production environment.

Acknowledgments

This research was supported in part by VIDAL (http://www.vidal.fr/) and the project FP7-ICT-1-5.2-Risk Assessment and Patient Safety (PSPIP). This work was supported in part by an appointment of A. Névéol to the NLM Research Participation Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education. 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., Mary V., Gaudinat A., Boyer, C., Rogozan, A., Darmoni SJ.: A Benchmark Evaluation of the French MeSH Indexing Systems. LNCS. Proc. AIME 2005, pp. 251-255.
4. 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.
5. 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.
6. 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.
7. Aronson AR, Mork JG, Gay CW, Humphrey SM, Rogers WJ. The NLM Indexing Initiative's Medical Text Indexer. Medinfo. 2004;2004:268-72.
8. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. Proc AMIA Symp. 2001:17-21.
9. Bodenreider O, Nelson SJ, Hole WT and Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp 1998:815-9.
10. 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.
11. 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].
12. 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].