Ubiquity: Technologies for Better Health in Aging Societies
A. Hasman et al. (Eds.)
IOS Press, 2006

A method of cross-lingual consumer health information retrieval

Aurélie NÉVÉOL[1, 2], Suzanne PEREIRA[1, 2, 3], Lina F. SOUALMIA[2] , Benoît THIRION[2], Stéfan J. DARMONI[1, 2]

[1] *Equipe GCSIS, Laboratoire PSI - INSA & Université de Rouen, France*
[2] *Equipe CISMeF, CHU de Rouen, France*
[3] *LERTIM, Faculté de Médecine de Marseille, France*

Abstract. Objectives: This paper presents a method of cross-language information retrieval aiming to make medical information available to patients in French and English, regardless of the query language they wish to use. Methods: We describe the two MeSH-related terminologies used in this work. We show that the French patient synonyms included in CISMeF can be automatically mapped to the English consumer-oriented health topics used in MEDLINEplus, via the MeSH thesaurus. The links between French and English patient terms thus inferred can subsequently be exploited to automatically translate patient queries. Results: 129 MEDLINEplus topics have been mapped to 142 CISMeF patient synonyms. Contextual links for cross-language retrieval have been added to the patient dedicated French information Gateway CISMeF. Conclusion: we have presented an efficient method for cross-lingual patient information retrieval in French and English, which may also be applied to other language pairs, subject to the availability of patient terminologies and of the MeSH thesaurus in these languages.
Keywords: Information Retrieval, MeSH, consumer health information, automatic translation

1. Introduction

An increasing amount of medical information is made available in electronic format, and published on the Internet. In recent years, the public has gained an easier access to these resources, and patients have also been encouraged to take advantage of this new media. The emergence of such technologies of information and communication in the medical domain calls for adequate tools to assist lay persons in their search for trustworthy health information intended for non specialist audiences, also known as Consumer Health Information (CHI).

Since 1995, CISMeF[1] (French acronym of Catalogue and Index of Medical On-Line Resources) has been selecting the most important resources of institutional health information in French [1]. It currently contains more than 15,000 resources intended for

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

health professionals (e.g. evidence-based resources, practice guidelines), medical students (e.g. lecture notes), and patients (e.g. patient education handouts). As in other health gateways (such as PubMed® or OMNI) the content description of resources referenced in CISMeF uses the reference thesaurus for indexing the bio-medical literature, namely MeSH® (Medical Subject Headings). However, patients –and even health professionals- often have doubts as to the proper MeSH denomination they should use to phrase their information query. Aware that inadequate queries may result in failure to obtain the desired information, the CISMeF team has addressed this issue in previous work by refining the information retrieval algorithm used in the catalogue's dedicated search engine Doc'CISMeF [2], [3] or enriching the list of MeSH synonyms available in French [4], [4], [6].

To provide further assistance to users, cross-language MeSH information retrieval is already available in CISMeF: MeSH queries can be entered in Doc'CISMeF either in French or in English. Besides, after the result of each query, contextual links to other trustworthy health gateways are provided to help users access health information in English. Therefore, a MeSH query entered in French in Doc'CISMeF will be translated into English, and a contextual link enables users to launch the translated query in PubMed or OMNI to access corresponding resources in English. However, similar work remains to be done for patient vernacular. In fact, if many patients have a sufficient grasp of the English language to benefit from health information in English, they experience difficulties in expressing their information need in a foreign language, and particularly with very specific terms as should be used in health gateways.

2. Objectives

This work aims to make medical information available to patients in several languages, regardless of the language they wish to use to state their information need. It is mainly intended for multilingual patients who may be able to understand medical information in several languages, but find it easier to phrase their information query in one particular language. In this paper, we describe a method of cross-language medical information retrieval in order to help patients with multilingual skills accessing health information in several languages. Specifically, we intend to use terminological resources in French and English to improve cross-lingual information retrieval in the CISMeF catalogue, and provide access to patient medical information in English in response to queries formulated in French.

3. Methods

3.1. CISMeF patient synonyms and MEDLINEplus topics

In the CISMeF terminology, CHI terms are defined as MeSH synonyms that patients are more likely to use than the actual MeSH terms. Therefore, more than one patient synonym may be available for a given MeSH term. For example, "tumeur osseuse" and "cancer des os" are two patient synonyms for the MeSH term "tumeurs des os" (bone neoplasms). On the other hand, CISMeF terminology experts are reluctant to use one particular term as a synonym for more than one MeSH term. These are very rare occurrences, and such synonyms are said to be ambiguous. This choice regarding the addition of entries in the terminology results in having a larger number of synonyms (N=531) than the number of MeSH terms impacted (N=431), as can be seen in table 1.

MEDLINEplus topics are health topics specifically selected for *consumers* – this includes patients, but also any lay person looking for reliable health information. The topics are meant to cover a large range of the consumers' health interests. For this reason, some topics may relate to more than one MeSH term. For example, the topic "AIDS" relates to the MeSH terms "Acquired Immunodeficiency Syndrome" and "HIV Infections". As a result, there are less health topics (N=698) than MeSH terms impacted (N=1130), as can be seen in table 1.

3.2. Linking CHI terms

CHI equivalents to MeSH terms have been developed independently in French and English by the CISMeF and MEDLINEplus teams. Thanks to the French version of the MeSH thesaurus developed by the French Medlar centers (INSERM), links between French and English MeSH terms are already available. The effort to enrich the MeSH with CHI terms in French (CISMeF patient-oriented synonyms) and English (MEDLINEplus topics) has led to the creation of semantic links between CHI terms and MeSH terms in each language. Figure 1 shows the links existing between the different types of terms involved.

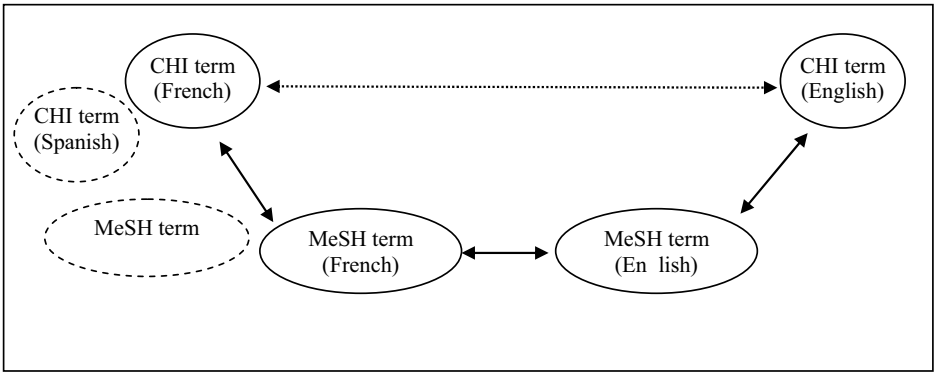

Figure 1: Method used for mapping French patient synonyms to English CHI equivalents

It is important to note that there are no explicit links between French and English CHI terms. However, the succession of three links (CHI to MeSH, MeSH to MeSH, and MeSH to CHI) makes it possible to induce semantic links between CHI in the different languages.

For example, linked to the English CHI term "second-hand smoking", we find the English MeSH term "tobacco smoke pollution", and its French MeSH equivalent, "pollution fumée tabac". There is one French CHI term linked to "pollution fume tabac", namely "tabagisme passif". Therefore, as shown on Figure 2, we can infer from these links that "tabagisme passif" is a French CHI equivalent for "second-hand smoking".

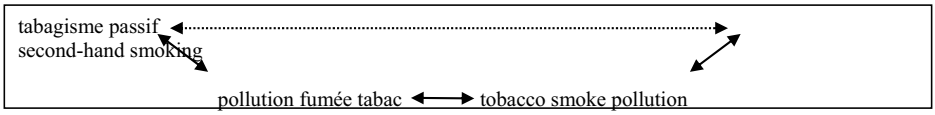

Figure 2: Sample mapping of CHI terms

4. Results

Table 1 presents the number of French (CISMeF) patient synonyms and English (MEDLINEplus) health topics used in our experiment, and the number of links that could be inferred between any pair of CHI terms. The second and third rows indicate the number and proportion of MeSH terms impacted (MeSH 2005 was the version used in this study).

	CHI (French)	**CHI (English)**	**Links created**
Rough number	531	698	280
MeSH impact (MeSH 2005)	431	1130	105
Proportion (MeSH 2005)	1.9%	4.9%	0.5%

Table 1: Number of Consumer specific terms in French and English, and links created

Table 2 provides examples of terms that could be mapped, and the links inferred. In cases where more than one CHI term was linked to a MeSH term in either French or English, more than one link could be inferred.

MeSH UID	**CHI (French)**	**CHI (English)**	**Links created**
D014028	tabagisme passif	Secondhand Smoke	tabagisme passif ↔ Secondhand

			Smoke
D001859	cancer des os tumeur osseuse	Bone Cancer	cancer des os ↔ Bone Cancer tumeur osseuse ↔ Bone Cancer
D014947	Trauma traumatisme	Injuries Wounds	trauma ↔ Injuries trauma ↔ Wounds traumatisme ↔ Injuries traumatisme ↔ Wounds

Table 2: Sample inferred links for three MeSH descriptors

Overall, 129 health topics were linked to 142 patient synonyms. As a result, 129 contextual links to MEDLINEplus topics pages will be created in French search engine Doc'CISMeF.

As an evaluation of the method, we can say that 18.5% of MEDLINEplus topics could be linked to at least one CISMeF patient synonym, and likewise, 26.7% of CISMeF patient synonyms could be linked to at least one MEDLINEplus topic.

5. Discussion

5.1. Links between CHI terms in French and English

Table 1 shows that, overall, the application of the method presented to infer links between French and English CHI terms covers only a small portion of the MeSH (0.5%). This is partly due to the limited MeSH coverage of the CHI terms (1.9% for French, 4.9% for English). In fact, the approach relies fully on the CHI information available in the terminologies it is applied to. In practice, this means that the coverage of links inferred will reflect the scope of CHI information in the terminologies used, and the overlap of CHI information between terminologies. However, one strong point of the method is that it is rather effortless as it exploits CHI information already available.

Even though the coverage of the links obtained for CHI terms in French and English is small on the MeSH scale, we consider these 129 sample links as representative of the performance of the method. All 129 links were reviewed and approved by a medical librarian.

Links such as those listed in Table 2 can be used in health search engines to answer patient queries. For example, when a patient enters the query "tabagisme passif", in addition to the list of CISMeF patient resources concerning "tabagisme passif ", French search engine Doc'CISMeF currently provides a dynamic link launching a MEDLINEplus search on "tobacco smoke pollution". Therefore, without knowing the scientific term referring to "tabagisme passif" in either French or English nor the translation of "tabagisme passif " in English, a patient can still access information on " tabagisme passif " both in French and English. In the near future, we are planning to exploit the CHI links inferred to make the dynamic links to MEDLINEplus more precise. As shown in Figure 3, a query on "tabagisme passif" entered in Doc'CISMeF will retrieve French resources on the subject

and provide a link to the MEDLINEplus page specifically dedicated to "secondhand smoke", which is the exact translation in English of "tabagisme passif".

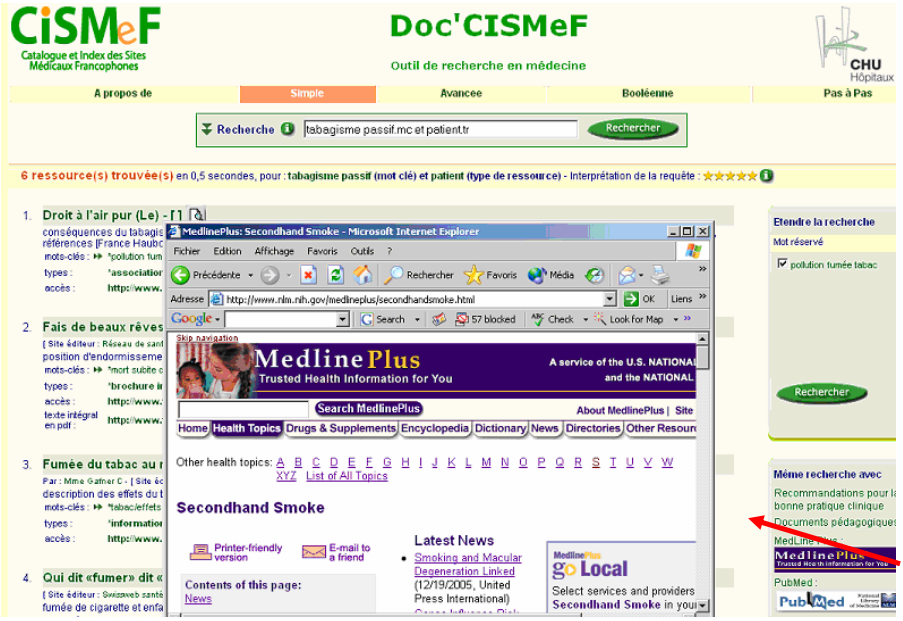

Figure 3: Prospective use of the inferred CHI links for Cross-Language Information Retrieval

5.2. Literature review

Thesauri-based Cross-Language Information Retrieval (CLIR). The problem we address, namely giving access to information in several languages for multilingual patients, answers a growing need already pointed out by Lu *et al.* [8] who developed a method of automatic translation of MeSH terms into Mandarin Chinese, and focused on the extraction of patient translations. In fact, the availability of many terminologies in the medical domain, including multi-lingual resources such as the MeSH paves the way for thesauri-based CLIR systems. A recent review of the CLIR literature [9] points out that the major drawback of thesauri-based CLIR methods is the availability of comprehensive multi-lingual thesauri. As these resources become available, thesauri-based CLIR should be preferable to query translation using either dictionaries or automatic translation software for which word-sense desambiguation in queries is difficult to overcome. A recent comparison [10] of thesaurus-based and automatic translation based CLIR systems show the superiority of the thesaurus (MeSH) based approach, but further work indicated that a combination of both methods could improve the performance.

Extending Thesauri-based CLIR. The application described above is the creation of contextual links to facilitate the interoperability between two different knowledge databases, in our example, CISMeF-patient and MEDLINEplus. However, Cimino *et al.* [11] have shown that this interoperability can be successfully extended to the creation of links between databases containing different types of information, such as a knowledge base and a patient record. In this respect, multi-lingual CHI terms can also be exploited to help patients understand their electronic files, and to search related information. In fact, while a patient is looking at his or her medical electronic record, multilingual information about the medical topics contained in the record could be made available through infobuttons. In 2005, a monolingual prototype application for health professional has been developed at the Rouen University Hospital [11], and we are planning to extend this functionality for patients and multilingual information retrieval.

Usability of contextual links for CLIR. A study of German consumer health search habits conducted in 2001 [7] showed that study participants with a good command of English did not try to search for information in this language. However, in spite of a substantial Internet experience (an average practice of 33 months), some of the participants seemed to lack understanding of search features, such as language selection in generic search tools or the use of domain dedicated engines and portals. Therefore, it seems important to inform users of the possibility for cross-language retrieval, and check whether its availability in the form of contextual links is suitable for their information needs. Previous work in the CISMeF team on contextual links to health information inserted in patient records for health professionals use received an informal positive feed-back, that will lead to business opportunities.

5.3. Perspectives

The method we used to infer links between CHI terms in French and English exploits the independent efforts of teams developing MeSH-related CHI vocabularies in various languages for which a MeSH translation is already available. The method is completely generic and can be applied to other language pairs such as English and Spanish in order to facilitate the access of bilingual patients to trustworthy consumer health information.

6. Conclusion

We have presented a method of cross-language medical information retrieval intended for patients, and we have applied it to a particular language pair, French and English. As a result, 280 links between French and English CHI terms have been inferred. The validation of the inferred links by an expert shows that the approach is relevant, and all the links will be used soon in the Doc'CISMeF search engine to launch precise cross-language patient queries.

Acknowledgements

This research was supported in part by an appointment of Prof. Darmoni 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 also like to thank Paula Kitendaugh at the National Library of Medicine in Bethesda, MD for granting them access to MEDLINEplus topics links to MeSH terms for this work and Gaëtan Kerdelhue at the University Hospital Library of Medicine in Rouen for his help in validating the CHI links.

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] Darmoni SJ, Thirion B, Platel S, Douyère M, Mourouga P, Leroy JP. - CISMeF-patient: A French counterpart to MEDLINEplus. JMLA Journal of the Medical Library Association 2002; April;90(2):248-53

[3] Soualmia LF, Darmoni SJ. Combining different standards and different approaches for health information retrieval in a quality-controlled gateway. Int J Med Inform. 2005 Mar;74(2-4):141-50.

[4] Névéol A., Ozdowska S.: Extraction de termes médicaux à partir d'un corpus parallèle anglais/français, Actes de Extraction et Gestion des Connaissances 2005:655-64

[5] Ozdowska S., Névéol A., Thirion, B.: Traduction compositionnelle automatique de bitermes dans des corpus anglais/français alignés. Actes des 6èmes rencontres Terminologie et Intelligence Artificielle, 2005:83-94.

[6] Deléger L., Merkel M., Zweigenbaum P. Using Word Alignment to Extend Multilingual Medical Terminologies. LREC workshop, 2006: to appear

[7] Eysenbach G and Köhler C: How do consumers search for and appraise health information on the world wide web? Qualitative study using focus grups, usability tests and in-depth interviews. BMJ 2002:324;573-577

[8] Lu WH, Lin SJ, Chan YC and Chen KH: Semi-automatic construction of the Chinese-English MeSH using web-based term translation method, Proc. AMIA Symp. 2005:475-9.

[9] Chiao YC: Extraction lexicale bilingue à partir de textes médicaux comparables – application à la recherche d'information crosslingue. PhD thesis, Université Paris VI. 2004.

[10] Ruch P:Query translation by Text Categorization, Proc. COLING 2004.

[11] Cimino JJ, Elhanan G, Zeng Q: Supporting infobuttons with terminological knowledge. Proc AMIA Annu Fall Symp 1997:528-32.

[12] Pereira, S: Etude de faisabilité de différentes méthodes d'optimisation du codage médico-économique. Master Thesis, Université Paris V:2005.