MEDINFO 2010
C. Safran et al. (Eds.)
IOS Press, 2010

doi:10.3233/978-1-60750-588-4-991

Characterizing Consumer Health Terminology in the Breast Cancer Field

Radja Messai[a,b], Michel Simonet[a], Nathalie Bricon-Souf[b], Mireille Mousseau[c]

[a] *TIMC-IMAG Laboratory, Joseph Fourier University, Grenoble, France*
[b] *CERIM, EA 2694, 1 Place de Verdun, 59045 Lille Cedex, France*
[c] *University Hospital of Grenoble, France*

Abstract

Despite the large availability of medical information on the Internet, health consumers still encounter problems to find, interpret and understand this information. These problems are mainly due to their lack in medical knowledge and the difference between their language and the language of health professionals. In order to propose information retrieval services more adapted to health consumers language and knowledge, we have developed techniques to collect, identify and analyze the terms and the expressions used by lay persons to talk about breast cancer. The study of health consumers' language is a relatively recent research field. Many studies have been conducted to analyze and characterize the vocabulary used by health consumers to talk about medical subjects in English. We have conducted the same study for the French language in the breast cancer field. We have gathered a corpus of texts to identify terms and expressions used by health consumers who talk about breast cancer in French. The terms have been organized in a concept-based terminology. This terminology has been analyzed on several levels: concept level, term level, term-concept level and finally relation level.

Keywords:

Consumer health terminology, Natural language processing, Breast cancer.

Introduction

A person when she/he is faced with a health problem needs to understand her/his illness, its diagnosis and the different treatment options that she/he is liable to follow. The doctor still remains the more natural way to find such information. However, more and more people are turning to different resources on health information.

A Eurobarometer survey published in 2003 [1] showed that the main source of health information used by European citizens is health professionals (pharmacists, doctors, etc.) 45.3%, followed by television 19.8%, books and medical encyclopaedia 7.7%. Internet is less widely used with 3.5%.

The survey also showed that only a small proportion (23.1%) of people in the EU use the Internet to find health information. Nevertheless, 41.5% of the people within the EU think that the Internet is a good way to get health information.

The Internet has become a popular way to access up-to-date information, and more and more people are turning to it to find information about health. However, technical, cultural and linguistic barriers are numerous when using Internet sites since most health-related information is available in English and uses specialized medical terminology [2-4].

A patient-oriented terminology aims at gathering the different ways lay people express themselves and talk about health topics, and linking them to the medical jargon used by health professionals. Using such a terminology will help bridging the communication gap between the two communities [5].

A terminology reflecting the patients' common language is the first phase of an ambitious project aiming at helping patients to better understand and master their health situation. We have worked on a common terminology in the field of breast cancer. This particular domain has been chosen because of its interest for the citizens and also as a testing ground for a methodology which could be extended both to other health domains and to other languages.

Although it has been long recognized that health consumers talk about and interpret differently medical concepts than health professionals, little efforts have been done to effectively build terminologies to bridge the gap between the two communities. These terminologies gather the terms used by health consumers to talk about their condition, and link them to the underlying medical terms and concepts.

An open source and collaborative development of a consumer health vocabulary has been initiated by the Harvard Medical School (HMS) and the National Library of Medicine in the USA [6]. Their aim is to develop an open source Consumer Health Vocabulary (CHV) by identifying consumer friendly names of medical concepts, correctly map them to UMLS (Unified Medical Language System) the reference terminology in the health domain, and create consumer health-specific concepts and relations. They have gathered 12 millions query log data and then performed automated term mapping and statistical analysis to select candidate terms for manual review. A web-based tool for collaborative review has also been devel-

oped. They have exhibited 90 000 concepts over the whole health domain in their consumer text corpus. The development of the CHV is still in progress.

In France, a team in the Rouen University Hospital has initiated the CISMeF project (acronym for Catalog and Index of French Language Health Resources on the Internet)[1] in February 1995. Its main objective is to catalog and index the most important and quality-controlled sources of institutional health information in French. For this purpose, CISMeF uses: the MeSH thesaurus (the National Library of Medicine's controlled vocabulary thesaurus) and several metadata element sets, including the Dublin Core (a metadata standard). In December 2007, the number of indexed resources totalled over 41 300 with a mean of 80 new resources each week.

CISMeF-patient is the French counterpart to MEDLINEplus [7]. It is a dedicated Website for patients, their families, and the general public. CISMeF-patient has been under development since 1997. CISMeF-patient has been created as a response to a growing need for consumer health information and to extend the awareness of quality health information resources available on the Internet. It uses the MeSH thesaurus end metadata to index web sites. The team has included many health consumer terms into the MeSH in order to facilitate health information seeking for non-professionals.

Materials and Methods

In this section, we describe the materials and the methodology used to develop the breast cancer terminology. This work is based on the experience of Tony Tse with the difference that T. Tse has worked on the whole health domain and for the English language [8].

Building a breast cancer terminology from a corpora of texts

A terminology is the set of words and expressions used to designate the concepts of a domain. A breast cancer terminology for lay people is made of the terms (i.e., words and expressions) that patients use to speak about breast cancer and also the terms they are liable to meet in their medical files or in the breast cancer literature. Therefore, a breast cancer terminology for lay people should contain terms specific to the patients' language, such as *breast pain* for *mastodynia*, but also medical terms such as *pyrexia*, which they are faced with.

Terms which are considered as synonyms are grouped into concepts. The concepts themselves are structured through different relationships. For example: *"Chemotherapy"* Is-A *"Breast cancer treatment"*. We have collected the terms from two types of corpus of texts: a mediator corpus and a health consumer corpus. Tony Tse calls "mediator corpus" a set of texts written by health information mediators (i.e., persons whose intent is to inform or influence the lay public about various medical topics, products, and services) and "health consumer corpus" texts written by participants in Web-based health discussion forums [8].

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

The mediator corpus has been built manually by selecting 575 documents issued from the answers of the search engine "Google" to the query "breast cancer". The selection has been done according to several criteria: domain representativeness, targeted public, page author, complexity of the used language. The consumer corpus has been built automatically by the extraction of 9 843 users' messages on two Web-based breast cancer discussion forums: The French League against Cancer and Essentielles.net.

We have used statistical methods to extract n-grams (a n-gram is a sequence of n consecutive words) from our corpora. We have obtained 6 896 candidate terms from the mediator corpus and 11 723 candidate terms form the consumer corpus.

The analysis of the list of candidate terms has been done manually with the help of a concondancer, a tool which helps visualizing each expression in its context [9]. It allows the user to look for terms in the corpus by using regular expressions and it produces concordances, (i.e., lists of occurrences of a term in a source text, surrounded by an appropriate portion of its original context). We have also studied the structure of web pages to identify the important concepts of the domain and to get a first hierarchy of concepts. The building of the terminology has been done progressively by studying every term and creating the appropriate concepts and relations every time it is needed.

The Protégé ontology editing tool has been used to represent the terminology in several standard languages including the W3C languages RDF(s) and OWL. By doing so, the terminology becomes usable by computer applications.

- We have tried to map the concepts of this terminology to those of UMLS and CHV. The connection between these terminologies has been done manually by using the UMLS identifiers, which are attached to the concepts in both terminologies (UMLS and CHV). Only the case of exact matches has been retained. We have obtained 83% of exact matching, 3% of partial correspondence and 14% of no correspondence.

Terminology analysis

The terminology has been analyzed on several levels:

- Term level;
- Concept level;
- Term-concept level;
- Relation level.

The objective of this analysis is to better understand the way lay persons talk about concepts and notions in the breast cancer field and structure them. Recent studies have shown significant differences between the professional and lay languages. However, these studies were conducted on the entire health domain and for the English language [8,10]. The produced terminology will be the core of an Information Retrieval system.

Term analysis

Many studies have used the length of the terms as an indicator of their complexity in order to evaluate the readability of documents [10-12]. We have compared the length of the terms coming from the two types of corpus. The results are shown in Table 1.

Table 1- Length of terms

	Health consumers	Mediators
Mean characters/term	21,5	22,8
Mean words/term	3.1	3

This comparison does not show significant differences between lay terms and mediators terms. The length of terms in this context is not an indicator of their complexity.

Concept analysis

The mapping of the breast cancer terminology to UMLS terms has revealed many interesting situations:

- Five concepts have multiple correspondences in UMLS. For example: *Cancer de l'ovaire* can be mapped to *Ovarian carcinoma* or *Malignant neoplasm of ovary.*
- Two pairs of concepts have a unique correspondence in UMLS. The concepts *Mammography* and *Mammogram* are mapped to the concept *Mammography* in UMLS. The same thing is observed for the concepts *Primipare (primipare)* and *Primiparité (primiparity)* and the concept *Primiparity.*

These cases show problems in the UMLS conceptualization. For example, *mammography* and *mammogram* designate two different concepts: a type of x-ray imaging used to create detailed images of the breast for the first, and an x-ray picture of the breast for the second.

Terms-Concepts analysis

Expressive variability of concepts: For each concept, we have calculated the expressive variability (number of terms which designate the concept) [13]. The objective of this step is to learn about the types of concepts with a high expressive variability. The mean of the expressive variability in the terminology is 2.16 terms. Most concepts are designated by one term (Figure 1).

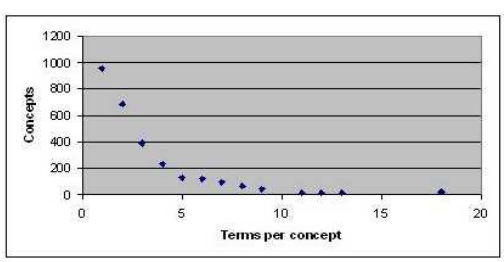

Figure 1- Terms distribution per concept

The study of the concepts with an expressive variability higher than 5 has shown that these concepts concerns medical concepts that we encounter in everyday life and which correspond to complex medical terms. Lay persons tend to describe concepts, which lead to a big production of terms. However, concepts with an expressive variability lower than 5 are of two types: either very well-known concepts like names of common organs (liver, lung, etc.) or highly technical terms like: tamoxifene.

Overlapping between health consumers and mediators terminology: We have compared the sets of terms coming from the two types of corpus (health consumers and mediators) in two steps:

1. Conceptual overlapping: identify the concepts common to both terminologies and the concepts specifics to each of them.
2. Terminological overlapping: for the common concepts, identify the terms common to both terminologies.

The Table 2 shows the results of this comparison.

Table 2- Overlapping between the two terminologies

	Common	Health consumers	mediators
Concepts	1 254	8	25
Terms	2 238	289	182

Relation analysis

Health consumers, in addition to their terminological problem, have often difficulties to understand how medical concepts are related. Among the defined relations in the terminology we have used the relation *Relation_X* to link two concepts without specifying the relation. This type of relation is used to define links between concepts that health consumers link without a medical argument. For example, the concept *contraceptive pill* is linked to the concept *breast cancer* because some health consumers believe that it is the case, although it is not scientifically established. *Relation_X* is also used to link the concepts which are not well understood by health consumers. For example, the concept *Vagina disorders* and the concept *Vaginitis* are linked by both relations *Is-A* and *Relation_X.* Most of health consumers believe that *vagnitis* embraces all the *vagina disorders*; however *vaginitis* represents only the inflammation of the woman's vagina. The use of this type of relation for representing this kind of phenomena preserves the "good" structure of the terminology.

Results

In the resulting breast cancer terminology, we have 1 287 concepts, designated by 2 783 terms in French. We have defined a set of 61 relations in addition to the classical Is-A and Part-Of relations to structure the concepts of the terminology.

Discussion and Conclusion

This work has shown some differences that exist between professional and health consumer terminologies. We have observed that the main differences are not at the concept level but at the term level. However, the current ontology representation languages do not offer the possibility to annotate a term by its "technical" level (lay or professional). An interesting alternative is offered by SKOS[2] which provides metadata to indicate the language of a term (i.e., lay or professional).

The bilingual terminology which has been built for breast cancer is the basis of future extensions to other health fields and other languages. The first considered application will be concept-based information retrieval, which will enable people to ask questions by using their everyday words and retrieve results in any language.

Such work is important for both patients and doctors because through a better understanding of her/his medical situation a patient will be able to better collaborate with the doctor, provide him more pertinent information on her/his situation and become a fully responsible partner in the decisions about her/his treatment. Informed Patients require less time for doctor explanations, and may be more likely to comply with doctors' instructions and to adopt a healthy lifestyle [14].

Acknowledgments

This work was supported by the French organizations: Ligue Contre le Cancer, Fédération Hospitalière de France and AGARO (Association Grenobloise d'Aide à la Recherche en Oncologie).

References

[1] EU survey results: Europe goes on-line for health information, but still prefers more traditional sources. EuropeMedia: April 18, 2003 issue; Ref. IP/03/550.

[2] McCray AT, Loane RF, Browne AC, et al. Terminology issues in user access to Web-based medical information. Proc AMIA Symp 1999:107-11.

[3] McCray AT, Tse T. Understanding search failures in consumer health information systems. Proc AMIA Symp 2003: 430-4.

[4] Zeng Q, Kogan S, Ash N, et al. Patient and clinician vocabulary: How different are they? Medinfo 2001;10(1):399-403.

[5] Plovnick RM, Zeng Q. Reformulation of consumer health queries with professional terminology: A pilot study. J Med Internet Res 2004;6(3):e27.

[6] Q. T. Zeng and T. Tse, A Case for Developing an Open-source First-Generation Consumer Health Vocabulary. J Am Med Inform Assoc, 2005.

[7] S. Darmoni, B. Thirion, S. Platel, M. Douyere, P. Mourouga, and J.-P. Leroy, "Cismefpatient: a french counterpart to medlineplus," J Med Libr Assoc, vol. 90, no. 2, pp. 248-253, 2002.

[8] Tse T. Identifying and Characterizing a Consumer Medical Vocabulary. PhD Thesis: College of Information Studies, University of Maryland, College Park; 2003.

[9] Bernhard D. Ontology Building Based on Text Corpora. Master Thesis: Institut National Polytechnique de Grenoble; 2003.

[10] Zeng Q, Kim H, Goryachev S, Keselman A, Slaughter L, Smith C. Text characteristics of clinical reports and their implications for the readability of personal health records. Stud Health Technol Inform. 2007; 129:1117-1121.

[11] Gemoets D, Rosemblat G, Tse T, Logan R. Assessing readability of consumer health information: an exploratory study. In: Medinfo. vol. 11; 2004. p. 869-873.

[12] Rosemblat G, Logan R, Tse T, Graham L. Text Features and Readability: Expert Evaluation of Consumer Health Text. In: Medical Internet. MEDNET; 2006.

[13] Tse T, Soergel D. Exploring Medical Expressions Used by Consumers and the Media: An Emerging View of Consumer Health Vocabularies. In: AMIA Annu Symp Proc; 2003. p. 674-678.

[14] Eysenbach G. The impact of the Internet on cancer outcomes. CA Cancer J Clin 2003; 53(6): 356-371.

Address for correspondence

Radja Messai
CERIM, EA 2694
1, Place de Verdun
59045 – Lille Cedex
France
radja.messai@univ-lille2.fr

[2] http://www.w3.org/2004/02/skos