A clinical problems terminology in French: extraction from free text sentences

Fabry P, Müller H, Lovis C

Service d'Informatique Médicale, Hôpitaux Universitaires de Genève, Suisse
Paul.Fabry@sim.hcuge.ch

Introduction

The Problem Oriented Medical Record (POMR) has been proposed since the late sixties in order to provide an efficient way to manage data within electronic medical records and has been promoted since by the American Institute of Medicine [1]. The POMR is arranged according to a list of clinical problems defined for the patient. Data are added in the record using progress notes indexed by problems. This problem list provides the functions of giving a brief, formal summary of the patient's medical history and of acting as an index for organizing the routine documentation produced by the patient's episode of care.

An effective implementation of the problem list in an electronic medical record calls for a controlled terminology of the problems. Several authors have evaluated the coverage of medical terminologies such as ICD-10, SNOMED, Read Codes or UMLS for this task [2, 3]. However, as these terminologies emphasize diagnoses or findings, they lack the completeness to express all categories of problems, and many institutions have developed their own local terminologies [4].

Another difficulty is the user acceptance of problem coding tools [5]. Physicians are not accustomed to limited vocabularies for expressing the patient's problems and may be reluctant to use a system which does not provide the "perfect" term. Problem lists, in opposition to the diagnostic list, must have a strong expressiveness of the loco-regional language habits.

Aim of this work is to test a method combining text processing operations and a statistics based term extraction algorithm in order to build up a problem terminology in French from free-text problem lists.

Materials and methods

Physicians enumerate clinical problems in admission notes, which are the first medical reports written at the patient's hospitalization.

Since 1993, the system in the University Hospitals of Geneva has been providing an electronic format to write theses notes. This format consists in a RTF document structured into five sections: reason for hospitalization, medical background, problems, diagnostic tests, and treatment.

In the "Problems" section each problem is noted down using free-text sentences. Physicians usually write admission notes in a "quick and dirty" way including many acronyms and abbreviations, wrong syntax and spelling errors. In addition to problem statements, sentences may also include other information such as test

results, planned treatment, etc.
We extracted a corpus of problem sentences from a set of randomly selected admission notes. The primary operation was to separate the problem statement from the rest of the sentence. We assumed that the problem statement was at the beginning of the sentence and we applied a method based on phrase boundary detection [6] which truncated each sentence starting at a set of specified words, nominal phrases or punctuation signs (e.g. "with", "context of", ":", etc.).
Then, the resulting problem statements underwent morphosyntactic (lemmatization, spell checking, stop word removal) and semantic (acronyms and abbreviations expansion, synonyms regroupment) normalizations.
We developed an algorithm that extracts word associations within the corpus using a mutual information metric [7].
Results were manually assessed and coverage is tested with a sample of randomly selected problem statements.

Results

This work has been restrained on a set of 5000 randomly selected medical admission notes collected in the University Hospitals of Geneva clinical system over 10 years. From these notes, we extracted 17 802 raw problem sentences. The mean of problems per admission note (± Standard Deviation) is 3.56 ± 1.84 with a range of 1 to 12.
The original corpus includes 126 717 occurrences of 17 466 different words, with a mean (± SD) of 7.1 ± 8.4 words by sentences. After truncation, morphosyntactic and semantic normalizations, a corpus of 58 075 occurrences of 3 210 different words is obtained with a mean (± SD) of 3.2 ± 2 words by problem statements.
Our method produced a total of 1 446 terms (Table 1).

Table 1. 5 most frequent terms and their occurrences.

Terms	Occurrences
Acute renal failure	550
Chronic obstructive pulmonary disease	625
High blood pressure	476
Chronic renal failure	439
Diabetes type 2	408

More than 88 % of these terms could be related to a relevant problem statement. We manually selected 998 terms for our problem terminology. We evaluated the coverage of this terminology with a set of 500 sentences randomly selected from other admission notes. For 383 (76.6 %), we found an appropriate problem statement in our terminology.
Several characteristics of the problem list can be outlined. First, problems include a large range of medical concepts: symptoms, diseases, abnormal tests results, care processes, etc. Although most terms have a precise medical meaning, some are rather ambivalent such as "care impossible at home". Finally, term granularity is extremely variable. There is, for example, just one term "social" for representing all

the social problems whereas there are almost a dozen for diabetes.

Discussion

Many tools already exist for term extraction from medical corpora in French, with statistical and morphosyntactic features [8, 9]. The approach proposed in this work is motivated by the low quality of textual information in admission notes, corrupted by misspelled words and conventional abbreviations.

Aim of this work is to produce a list of the most frequent terms used for describing problems. Resulting terminology is not exhaustive and should be considered as an "open" list that will be enhanced.

Expected benefit of such a list is to facilitate information retrieval within a electronic medical record. Other existing terminology, such as MeSH, could be coupled with a problem terminology in order to support bibliographical research from patient records.

Acknowledgement

This work has been funded by the Swiss National Science Foundation (SNF 632-066041).

References

[1] Dick R, Steen E, editors. The computer based patient record: An essential technology for health care. Washington DC: National Academy Press; 1991.

[2] Campbell JR, Payne TH. A comparison of four schemes for codification of problem lists. Proc Annu Symp Comput Appl Med Care 1994:201-5.

[3] Goldberg H, Goldsmith D, Law V, Keck K, Tuttle M, Safran C. An evaluation of UMLS as a controlled terminology for the Problem List Toolkit. Medinfo 1998;9 Pt 1:609-12.

[4] Brown SH, Miller RA, Camp HN, Guise DA, Walker HK. Empirical derivation of an electronic clinically useful problem statement system. Ann Intern Med 1999;131(2):117-26.

[5] Wang SJ, Bates DW, Chueh HC, Karson AS, Maviglia SM, Greim JA, et al. Automated coded ambulatory problem lists: evaluation of a vocabulary and a data entry tool. Int J Med Inf 2003;72(1-3):17-28.

[6] Herzig T, Johns M. Extraction of medical information from textual sources: a statistical variant of the boundary word method. Proc AMIA Annu Fall Symp 1997.

[7] Church K, Hanks P. Word association Norms, Mutual Information, and Lexicography. Computational Linguistics 1990;16(1):22-9.

[8] Le Moigno S, Charlet J, Bourigault D, Degoulet P, Jaulent MC. Terminology extraction from text to build an ontology in surgical intensive care. Proc AMIA Symp 2002:430-4.

[9] Baud RH, Lovis C, Rassinoux AM, Scherrer JR. Morpho-semantic parsing of medical expressions. Proc AMIA Symp 1998:760-4.