Identifying Proper Names in Parallel Medical Terminologies

Olivier Bodenreider[1], Pierre Zweigenbaum[2]

(1) U.S. National Library of Medicine, Bethesda, MD, USA
(2) DIAM – Service d'Informatique Médicale, DSI, AP-HP
and Département de Biomathématiques, Université Paris 6, Paris, France

We propose several criteria to identify proper names in biomedical terminologies. Traditional, pattern-based methods that rely on the immediate context of a proper name are not applicable. However, the availability of translations of some terminologies supports methods based on invariant words instead. A combination of five criteria achieved 86% precision and 88% recall on the 16,401 word forms of the International Classification of Diseases.

1 Introduction

Specialized domains normally have an extensive technical terminology associated with them. The vast majority of terms occurring in such a vocabulary are noun phrases in which the head noun is modified either by an adjective or a prepositional phrase. Proper names also play a role in the construction of complex noun phrases, and biomedical discourse is particularly rich in this phenomenon, for example "Parkinson's disease," "Achilles tendon" and "pouch of Douglas." Although the Unified Medical Language System® (UMLS®) [1] includes a list of proper names, it is far from complete. An effective, comprehensive, electronic lexicon of biomedical proper names is not currently available.

A reliable list of proper names would be a useful resource for a lexical tool such as the lexical variant generation program included in the UMLS. It would prevent some lexical variants from being computed for words not included in the lexicon (such as proper names), and would enhance spelling correction capabilities. Integrated into the resources of a terminology server, this would allow better mapping to the corresponding concepts. In the analysis of discharge summaries, for example, such a list would help distinguish between proper names from medical terms and the names of patients and doctors, hospitals and wards, cities and countries.

The present work is aimed at designing methods to collect proper names used in biomedical terminology, such as the International Classification of Diseases (ICD). The problem addressed here is somewhat different from that of the named entity recognition task as defined in the Information Extraction community [2]. In that task, expressions containing proper names must be identified in running text and classified as referring to persons, organizations or geographical locations. Here, the task is to recognize that a word is a proper name.

Stud Health Technol Inform 2000;77:443-447.

Various methods for the identification, semantic categorization and disambiguation of proper names have appeared in the literature [e.g. 3, 4, 5]. These methods, however, are not always applicable to medical terminologies:

- Syntactic analysis might not be effective since medical terms are often limited to a few words.
- Medical proper names are essentially names of persons, with no apparent regular morphological properties.
- Capitalization is not always available and cannot be considered a definitive clue, since, for example, the first word of a term is capitalized in some terminologies.
- Lexical markers, found in the vicinity of proper names, are extremely variable in the biomedical domain. Virtually any body part ("Lisfranc's tubercle", "Gasser's ganglion"), condition ("Cheynes-Stokes respiration", "Graham Steell's murmur") or procedure ("Hartmann's colostomy", "Heimlich maneuver"), to name but a few types, can be named after some famous scientist.
- In addition, context is limited to its simplest form, the term itself; no further external clues are available.

Some external help, however, is available in that international medical terminologies have translations in several languages. Each "concept" of the terminology has a unique identifier which is shared by equivalent terms in different languages. This parallelism has already been exploited by Baud to collect multilingual medical dictionaries [6]. For example, the French ICD-10 has terms "maladie de Parkinson", "respiration de Cheynes-Stokes", "souffle de Graham Steell." Proper names are often spelled identically across languages, so that they generally belong to the set of invariant words of ICD-10 translation pairs. Parallel terminologies therefore provide us with a supplemental clue to detect proper names in terms.

2 Background

2.1 The International Classification of Diseases

The International Classification of Diseases has been developed and enhanced for more than a century. It is widely used throughout the world to record causes of death and to classify diseases, injuries and related health problems. Currently, the World Health Organization is in charge of its maintenance. The 10th revision (ICD-10) has been published in English in 1992. Translations of ICD-10 are now available in more than 20 languages [7].

ICD-10 consists of several volumes including a tabular list of some 17,000 diseases and an alphabetical index for the diseases (42,000 entries in the French version). Terms from the tabular list are usually noun phrases of variable length (Table 1). The syntactic structure of the index terms, however, is not as well defined. In the context of a parent term located above, only the modifiers of the term are mentioned, the other words of the term being replaced by hyphens to represent the depth. Examples of terms found in the alphabetical index under "Glaucoma" are given in Table 2. Here, "— — noncongestive" actually means "noncongestive chronic glaucoma." Most of the terms are followed by a code referring to diseases in the tabular list.

The major source of proper names in ICD-10 is the alphabetical index. It reflects the usage of terms (from both a lexical and a cultural point of view) in a particular language, especially of eponymic terms, whereas the tabular list represents a set of terms whose

structure is quite similar across languages. For this reason, proper names found in one translation of the alphabetical index of ICD-10 are not necessarily present in other translations. Finally, compound proper names found in different translations of ICD-10 sometimes show variation in the order or even in the number of the names (e.g. an alternate term for "relapsing panniculitis" is "Weber-Christian disease" in English, "maladie de Weber-Christian" in French, but "Pfeifer-Weber-Christian-Krankheit" in German).

Table 1. Examples of ICD-10 terms (tabular list).

C46.0	Kaposi's sarcoma of skin
E24	Cushing's syndrome
Q15.0	Congenital glaucoma
R51	Headache
T44	Poisoning by drugs primarily affecting the autonomic nervous system

Table 2. Examples of ICD-10 terms (alphabetical index).

Glaucoma	H40.9
— with pseudoexfoliation of lens	H40.1
— acute	H40.2
— chronic	H40.1
— — noncongestive	H40.1
— congenital	Q15.0

2.2 The reference list of proper names

The reference list of proper names used as our gold standard was established as follows. The vocabulary extracted from the French ICD-10 was compared to a large French lexicon augmented with medical terms. Words unknown from the lexicon were reviewed manually and assigned a syntactic category (NPR for proper names). Out of the 16,401 words from the French ICD-10, 1,212 are proper names.

3 Methods

We have studied several criteria that help classify words as proper names in biomedical terminologies. Since no individual criterion achieves both high precision and high recall, we have defined a combination of criteria that together support effective identification of the target words.

3.1 Individual criteria

Invariant words (INV). Words common to several translations of ICD are either generally proper names or foreign terms (including Latin). This criterion relies on the assumption that the majority of proper names are invariant words (i.e. their spelling does not change across languages), and most invariant words are proper names. Terms from three translations of ICD-10 (English, French and German) were tokenized. The list of invariant words was then defined by the intersection of the lists of words for each language. Due to differences in spelling or in the use of some proper names across languages, we investigated two types of criteria based on invariant words. The first one (INV3) defines a word as an invariant if it is found in each of the three translations. The second version (INV2) is more permissive and requires that the word be found in at least two of the three languages. Compared to INV3, INV2 finds twice as many invariant words, but only 15% more proper names.

Capitalization (CAP). Proper names usually start with an uppercase character. Since ICD makes consistent use of capitalization, this criterion is helpful in identifying proper names. Words other than proper names, however, are also capitalized, such as the first word of a term, making it impossible to use this criterion alone (e.g. in the term "Dementia in Parkinson's disease", "Dementia" is capitalized, yet is not a proper name).

The following two criteria deal with other types of capitalized word forms.

Symbols (N-S). Although capitalized, acronyms (all caps) and various abbreviations or symbols (also containing numerals) can be easily filtered out from proper names (e.g. “AIDS”, “[vitamin] B12”, “46,XX [karyotype]”). A character length of 2 or less is also used to filter out symbols from a list of candidates.

Micro-organisms (N-M). By convention, in micro-organism names, the first part referring to the genus is capitalized whereas the second part (species) is not (e.g. “Haemophilus influenzae”). Micro-organisms, however, can be easily identified by using a list of their names. Such lists, although extensive and not included in ICD itself, are not difficult to acquire from public domain repositories of biomedical taxonomies.

Patterns (PAT). We have shown earlier that proper names can be introduced by a large variety of words in biomedical terms. In the French ICD-10 only, we have identified 154 possible contexts. Therefore, the pattern is solely based on the prepositional construct from which eponyms are formed in French, and on the morphologic characteristics of compound proper names (e.g. “maladie *de* Parkinson”, “syndrome *d’*Albers-Schönberg”). Following an ICD convention, the right context is searched for instead when terms come from the alphabetical index (e.g. “Cooley, anémie ou maladie *de*”).

3.2 Combination

Some combination of the aforementioned criteria is necessary to achieve both high precision and high recall. The three major criteria are capitalization, invariant words and patterns.

In order to increase precision, additional criteria can be associated with any of the major criteria to take advantage of typographic conventions for proper names (CAP), morphologic characteristics of acronyms and symbols (N-S) or external resources such as a list of micro-organism names (N-M). Finally, even when refined, these criteria must be combined in order to increase recall. Such a combination would identify as a proper name any word which either occurred invariant or in one of the patterns, with additional filtering obtained by capitalization-based criteria in order to improve precision.

C-CAP defines a sort of baseline by filtering out uncapitalized words, acronyms and symbols, and micro-organism names. C-INV identifies proper names from the invariant words, and C-PAT applies a pattern to the terms. In both cases (C-INV or C-PAT), additional filtering is provided by associating the criteria defined in C-CAP. COMBO realizes the union of the proper names identified by two methods, one based on invariant word (C-INV) and one based on patterns (C-PAT), as shown in Table 3.

Table 3. Combination of individual criteria.

(C-CAP)							CAP	and	N-S	and	N-M
(C-INV)		INV				and	CAP	and	N-S	and	N-M
(C-PAT)				PAT		and	CAP	and	N-S	and	N-M
(COMBO)	(	INV	or	PAT	)	and	CAP	and	N-S	and	N-M

4 Results

Standard precision and recall values have been computed for each criterion (Table 4) or combination thereof (Table 5), as defined in the Methods section, on the 16,401 word forms of ICD-10, against the reference list of proper names.

C-INV defines the combination of criteria refining any method based on invariant words, INV2 or INV3. COMBO defines the reunion of C-INV and C-PAT, using either INV2 (for COMBO2) or INV3 (for COMBO3). COMBO3 represents the best tradeoff with both high precision and high recall, whereas C-PAT obtains the highest precision with a somewhat lower recall.

Table 4. Performance of individual criteria.

Criterion	INV2	INV3	CAP	N-S	N-M	PAT
Precision	0,289	0,473	0,276	0,075	0,073	0,482
Recall	0,874	0,713	0,993	1,000	0,950	0,837

Table 5. Performance of combined criteria.

Criterion	C-INV2	C-INV3	C-PAT	C-CAP	COMBO2	COMBO3
Precision	0,663	0,827	0,991	0,280	0,683	0,859
Recall	0,822	0,666	0,807	0,944	0,909	0,881

5 Discussion

The benefit expected from this work is to limit the human effort necessary to build a list of proper names used in medical discourse. Although imperfect, the performance of the combined methods is sufficient to reach this goal. Furthermore, the method proposed is both simple (it relies on basic lexical techniques) and robust (it does not rely on a strict alignment of the terminologies).

Different combinations of criteria could be applied to fit the characteristics of other medical terminologies available in several languages. Optimal combinations of criteria could be sought using statistical techniques. Strategies for the analysis of a new terminology could then be inferred by comparing the characteristics of the terminology to the performance of the criteria obtained in previous studies.

Experiments conducted on the SNOMED Microglossary for Pathology have confirmed the results obtained on ICD. We plan to test the strategic hypotheses on the Medical Subject Headings (MeSH).

6 References

[1] UMLS. UMLS Knowledge Sources. 10th ed. Bethesda (MD): National Library of Medicine; 1999.

[2] MUC-6. Proceedings of the Sixth Message Understanding Conference (Columbia, Maryland, Nov 1995): Morgan Kaufmann; 1996.

[3] D. McDonald. Internal and external evidence in the identification and semantic categorization of proper names. In: Boguraev B, Pustejovsky J, editors. Corpus Processing for Lexical Acquisition. Cambridge, Mass.: MIT Press; 1993. p. 61-76.

[4] C. Thielen. An Approach to Proper Name Tagging for German. In: From Texts to Tags: Issues in Multilingual Language Analysis - Proceedings of the ACL SIGDAT Workshop, Dublin; 1995. p. 35-40.

[5] N. Wacholder, Y. Ravin, M. Choi. Disambiguating proper names in text. In: Proceedings of the Applied Natural Language Processing, Washington, DC; 1997. p. 202-208.

[6] R. Baud, C. Lovis, A.-M. Rassinoux, P.-A. Michel, J.-R. Scherrer. Automatic extraction of linguistic knowledge from an international classification. Medinfo 1998;9(Pt 1):581-5.

[7] ICD-10. International Statistical Classification of Diseases and Related Health Problems. Geneva (Switzerland): World Health Organization; 1992.