Lexically-Based Terminology Structuring: Some Inherent Limits

Natalia Grabar and **Pierre Zweigenbaum**
STIM/DSI, Assistance Publique – Hôpitaux de Paris
& Département de Biomathématiques, Université Paris 6
{ngr,pz}@biomath.jussieu.fr
http://www.biomath.jussieu.fr/~{ngr,pz}

Abstract

Terminology structuring has been the subject of much work in the context of terms extracted from corpora: given a set of terms, obtained from an existing resource or extracted from a corpus, identifying hierarchical (or other types of) relations between these terms. The present paper focusses on terminology structuring by lexical methods, which match terms on the basis on their content words, taking morphological variants into account. Experiments are done on a 'flat' list of terms obtained from an originally hierarchically-structured terminology: the French version of the US National Library of Medicine MeSH thesaurus. We compare the lexically-induced relations with the original MeSH relations: after a quantitative evaluation of their congruence through recall and precision metrics, we perform a qualitative, human analysis of the 'new' relations not present in the MeSH. This analysis shows, on the one hand, the limits of the lexical structuring method. On the other hand, it also reveals some specific structuring choices and naming conventions made by the MeSH designers, and emphasizes ontological commitments that cannot be left to automatic structuring.

1 Background

Terminology structuring, *i.e.*, organizing a set of terms through semantic relations, is one of the difficult issues that have to be addressed when building terminological resources. These relations include subsumption or hyperonymy (the *is-a* relation), meronymy (*part-of* and its variants), as well as other, diverse relations, sometimes called 'transversal' (*e.g.*, *cause*, or the general *see also*).

Various methods have been proposed to discover relations between terms (see Jacquemin and Bourigault (2002) for a review). We divide them into *internal* and *external* methods, in the same way as McDonald (1993) for proper names. Internal methods look at the constituency of terms, and compare terms based on the words they contain. Term matching can rely directly on raw word forms (Bodenreider et al., 2001), on morphological variants (Jacquemin and Tzoukermann, 1999), on syntactic structure (Bourigault, 1994; Jacquemin and Tzoukermann, 1999) or on semantic variants (synonyms, hyperonyms, etc.) (Hamon et al., 1998). External methods take advantage of the context in which terms occur: they examine the behavior of terms in corpora. Distributional methods group terms that occur in similar contexts (Grefenstette, 1994). The detection of appropriate syntactic patterns of cooccurrence is another method to uncover relations between terms in corpora (Hearst, 1992; Séguéla and Aussenac, 1999).

In previous work we applied lexical methods to identify relations between terms on the basis on their content words, taking morphological variants into account. Our goal was then to assess the feasibility of such structuring by studying it on an existing, hierarchically structured terminology. Ignoring this existing structure and starting from the set of its terms, we attempt to discover hierarchical term-to-term links and compare them with the preexisting relations.

Our goal in the present paper is to analyze 'new' relations. 'New' means that these induced relations are not present in the original hierarchical structure of the MeSH thesaurus; they might nevertheless reflect useful links. Performing this analysis allows us to propose a more precise evaluation of the methods and their results and to point out some inherent limits.

After the exposition of the data we used in our experiments (section 2), we present methods (section 3) for generating hierarchical links between terms through the study of lexical inclusion and for

evaluating their quality with appropriate recall and precision metrics. We then present the analysis of some 'new' induced relations and attempt to propose a typology of term dependency in these relations (section 4). We finally discuss the limits of lexical methods for the structuring task (section 5).

2 The MeSH biomedical thesaurus, and associated morphological knowledge

We first present the existing hierarchically structured thesaurus, a 'stop word' list and morphological knowledge involved in the present work.

2.1 The MeSH biomedical thesaurus

The Medical Subject Headings (MeSH, NLM (2001a)) is one of the main international medical terminologies (see, *e.g.*, Cimino (1996) for a presentation of medical terminologies). It is a thesaurus specifically designed for information retrieval in the biomedical domain. The MeSH is used to index the international biomedical literature in the Medline bibliographic database. The French version of the MeSH (INSERM, 2000) contains a translation of these terms (19,638 terms) plus synonyms. It happens to be written in unaccented, uppercase letters. Both the American and French MeSH can be found in the UMLS Metathesaurus (NLM, 2001b), which can be obtained through a convention with the National Library of Medicine.

The concept names (*main headings*) which the MeSH contains have been designed to reflect their broad meanings and to facilitate their use by human indexers and librarians. In that, they follow a tradition in information sciences, and are not necessarily the expressions used in naturally occurring biomedical documents. The MeSH can be considered as a fine-grained thesaurus: concepts are chosen to insure a good coverage of the biomedical domain (Zweigenbaum, 1999).

As many other medical terminologies, the MeSH has a hierarchical structure: 'narrower' concepts (children) are related to 'broader' concepts (parents). This both covers the usual *is-a* relation and partitive relations (*part-of*, *conceptual-part-of* and *process-of*). The MeSH also includes *see-also* relations, which we do not take into account in the present experiments. This structure has also been designed in the aim to be intellectually accessible to users: an indexer must be able to assign a given concept to an article and a clinician must be able to find a given concept in the tree hierarchy (Nelson et al., 2001). To conclude, the MeSH team aims to organize it in a clear and intuitive manner, both for concept naming and concept placement.

The version of the French MeSH we used in these experiments contains 19,638 terms, 26,094 direct child-to-parent links and (under transitive closure) 95,815 direct or indirect child-to-ancestor links.

2.2 Stop word list

The aim of using a 'stop word' list is to remove from term comparison very frequent words which are considered not to be content-bearing, hence 'non-significant' for terminology structuring. We used in this experiment a short stop word list (15 word forms). It contains the few frequent grammatical words, such as articles and prepositions, that occur in MeSH terms.

2.3 Morphological knowledge

The morphological knowledge involved consists of *lemma*/*derived-word* or *lemma*/*inflected form* pairs where the first is the 'normalized' form and the second a 'variant' form.

Inflection produces the various forms of a given word such as plural, feminine or the multiple forms of a verb according to person, tense, etc.: *intervention – interventions*, *acid – acids*. We perform the reverse process (lemmatization), reducing an inflected form to its lemma (canonical form). We worked with two alternate lexicons. The first one is based on a general French lexicon (ABU, abu.cnam.fr/DICO) which we have augmented with pairs obtained from medical corpora processed through a tagger/lemmatizer (in cardiology, hematology, intensive care, and drug monographs): it totals 219,759 pairs (where the inflected form is different from the lemma). The second lexicon, more specialized and tuned to the vocabulary in medical terminologies, is the result of applying rules acquired in previous work from two other medical terminologies (ICD-10 and SNOMED) to the vocabulary in the MeSH, ICD-10 and SNOMED (total: 2,889 pairs).

Derivation produces, *e.g.*, the adjectival form of a noun (noun *aorta* $\leftrightarrow$ adjective *aortic*), the nominal form of a verb (verb *intervene* $\leftrightarrow$ noun *intervention*), or the adverbial form of an adjective (adjective *human* $\leftrightarrow$ adverb *humanely*). We perform linguistically-motivated stemming to reduce a derived word to its base word. For derivation, we also used resources acquired in previous work which, once combined with inflection pairs, results in 4,517 pairs.

Compounding, which combines several radicals, often of Greek or Latin origin, to obtain complex words (*e.g.*, *aorta* + *coronary* yields *aortocoronary*), has not been used because we do not have a reliable procedure to segment a compound into its component morphemes.

3 Acquiring links through lexical inclusion of terms

The present work induces hierarchical relations between terms when the constituent words of one term lexically include those of the second term (section 3.1). When comparing these relations with those that preexist in the MeSH, precision can reach 29.3% and recall 13.7% (section 3.2). We focus here on the analysis of the relations that are not found in the MeSH (section 3.3), which we develop in the next section (section 4).

3.1 Lexical inclusion

The method we use here for inducing hierarchical relations between terms is basically a test of *lexical inclusion*: we check whether a term P (*parent*) is 'included' in another term C (*child*), *i.e.*, whether all words in P occur in C. We assume that this type of inclusion is a clue of a hierarchical relation between terms, as in *acides gras* / *acides gras indispensables* (*fatty acids* / *fatty acids, essential*).

To detect this type of relation, we test whether all the content words of P occur in C. We do this on segmented terms with a gradually increasing normalization on word forms. Basic normalizations are performed first: conversion to lower case, removal of punctuation, of numbers and of 'stop words'. Subsequent normalizations rely on morphological ressources: lemmatization (with the two alternate inflectional lexicons) and stemming with a derivational lexicon. Terms are indexed by their words to speed up the computation of term inclusion over all term pairs of the whole MeSH thesaurus.

3.2 Application to MeSH and quantification

This structuring method has been applied to the flat list of 19,638 terms of the MeSH thesaurus. As expected, the number of links induced between terms increases when applying inflectional normalization and again with derivational normalization.

We evaluated the quality of the links obtained with this approach by comparing them automatically with the original structure of the MeSH and computing recall and precision metrics. We summarize here the main results; a detailed evaluation can be found in (Grabar and Zweigenbaum, 2002).

Depending on the normalization, up to 29.3% of the links found are correct (precision), and up to 13.7% of the direct MeSH links are found by lexical inclusion (recall). We also examined whether each term was correctly placed under one of its ancestors: this was true for up to 26% of the terms (recall); and the placement advices were correct in up to 58% of the cases (precision). The recall of links increases when applying more complete morphological knowledge (inflection then derivation). The evolution of precision is opposite: injection of more extensive morphological knowledge (derivation *vs* inflection) leads to taking more 'chances' for generating links between terms: the precision with no normalization (*raw* results) is 29.3% *vs* 22.5% when using all normalizations (*lem-stem-med*). Depending on the type of normalization, the best precision obtained for links is 43%.

3.3 Human analysis of 'new' relations

The evaluations presented in the previous section quantify the match between the induced relations and existing MeSH relations. However, they give no explanation for the fact that 70% of the induced relations are not considered relevant by the MeSH. This is what we study in the remainder of this paper: why these terms are not hierarchically related in the MeSH, and what kinds of relations exist between them.

According to the position of the words of the 'parent' term in the 'child' term, we divide the extra-MeSH relations into three sets: (1) the parent concept is at the *head* position in the child concept: *absorption*/*absorption intestinale*; (2) the parent concept is at the tail (*expansion*) position in the child concept: *abdomen*/*tumeur abdomen*; (3) *other* types of positions. Each set of relations is sampled by randomly selecting a 20% subset, both without normalization (*raw*) and with inflectional and derivational normalizations (*med-lem-stem*). Table 1 presents the number of analyzed relations (total = 194).

Normalizations	Head	Expan.	Other
raw	22	31	14
lem-stem-med	37	57	33

Table 1: Relations to analyze: sample sizes.

4 An analysis of new, lexically-induced relations

We first examine the issues encountered when trying to identify the head of each term (section 4.1), then review in turn each analyzed subset: head (section 4.2), expansion (section 4.3) and other relations (section 4.4).

4.1 Finding the head

In French, the semantic head of a noun phrase is usually located at the beginning of this phrase (this contrasts with English, where the semantic head is generally at the end of NPs). Moreover, as is often the case with terms, MeSH terms do not include determiners, so that the semantic head is usually the first word here. We therefore rely on a heuristic for determining 'head' and 'expansion' subsets: the head is the first word of the term, and the expansion is the last word. This is correct most of the time, but in some cases, the semantic head is positioned at the end of the term, generally separated with a comma, a tradition sometimes followed in thesauri:

filoviridae/*filoviridae, infections*,
leishmania/*leishmania tropica, infection*,
quinones/*quinone reductases*,
neurone/*neurone moteur, maladie*,
syndrome/*bouche main pied, syndrome*.

These cases must be hand-corrected and distributed into the following classes.

We also encountered another kind of error, due to overzealous derivational knowledge:

contracture/*contraction musculaire*,
biologie/*testament biologique*,

where *contracture* (a muscle disease) and *contraction* (normal muscle function) have both been stemmed to the same base word; the expansion adjective *biologique* is derived from the noun *biologie*, but its sense is generally more specific than *biologie*.

4.2 'Head' subset

Let us first discard a case where it seems that we encountered a translation error. An examination of the structure of the English MeSH and a search on Web pages show that in the French MeSH, *acide linoleique alpha* should read *acide linolenique alpha*, which is a kind of *acide linolenique* (and not a kind of *acide linoleique*). The induced relation:

acide linoleique/*acide linoleique alpha*

is therefore incorrect; with the correct spelling, the lexical inclusion:

acide linolenique/*acide linolenique alpha*

would reveal a correct hierarchical relation.

4.2.1 The head is not the 'genus' of the term

We encountered cases where the whole term did not have an *is-a* relation with the head as defined above. This happens in two types of situations.

The first situation is due to syntactic reasons. In the following induced relation,

acides amines / *acides amines, peptides et proteines*,

the larger term is an enumeration, with the sense of a logical OR. It is therefore the genus term, of which each of its components (*e.g.*, *acides amines*) is a sub-type.

The second situation is due to semantic reasons. Lexical induction of hierarchical relations assumes inheritance of the defining features of the genus term (*e.g.*, a *fatty acid, essential* is a kind of *fatty acid*). However, it is well known that this is not always true: a *plaster cat* is not a *cat* (*i.e.*, a mammal, etc.). This is sometimes modeled as a type coercion phenomenon. We found quite a few 'plaster cats' in our terms:

personnalite/*personnalite compulsive*,
voix/*voix oesophagienne*.

For instance, *personnalite* here describes 'behavior-response patterns that characterize the individual', whereas *personnalite compulsive* (*compulsive personality disorder*) describes a mental disorder. Disorders (or diseases) are different objects than behaviors in the MeSH.

4.2.2 The head is ambiguous

This depends on the choice of term names in the terminology (here, the MeSH). Terms like *absorption*, *investissement*, etc., have specific senses that make them polysemous. To determine a precise sense, these terms have to be specialized by their contexts:

investissement/*investissement (psychanalyse)*,
absorption/*absorption cutanee*,
goitre/*goitre ovarien*

Here, *investissement* alone (*investment*) has the financial sense, whereas in *investissement (psychanalyse)*, it has its more generic sense. In a similar way, *absorption* has a specific meaning in chemistry, and *goitre* alone is a disorder of the thyroid

gland. These cases are often non-ambiguous in the original English version of the same terms: for instance, *investissement (psychanalyse)* (fr) is a translation of *cathexis* (en).

A related case occurs when the name of a parent term is underspecified:

acides/*acides pentanoiques*,
acne/*acne rosacee*.

In these examples, *acides* means *inorganic acids*[1] and *acne* means *acne vulgaris*, but the convention adopted is to use these single words to name the corresponding concepts.

4.2.3 Ontological commitment

Finally, some induced links, although absent from the MeSH, are potentially correct *is-a* links, but the designers of the MeSH have made a different modeling choice:

amyotrophies/*amyotrophies spinales enfance*,
hyperplasie/*hyperplasie epitheliale focale*,
centre public sante/*centre public sante mentale*,
rectocolite/*rectocolite hemorragique*,
penicillines/*penicilline g*.

A general representational choice in the MeSH, as in some other medical terminologies (*e.g.*, SNOMED), is to differentiate on the one hand *signs or symptoms* and on the other hand *diseases* (a more fully characterized pathological state). This is the case for *amyotrophies* and *hyperplasie* (*signs or symptoms*) *vs amyotrophies spinales enfance* and *hyperplasie epitheliale focale* (*disease* of the nervous system, of the mouth).

For some reason, a *centre public sante mentale* (public mental health center) is considered not to share all the attributes of a general *centre public sante* (public health center), which prevents them from being in a parent-child relationship: they are only siblings in the MeSH thesaurus.

Penicillines, in the MeSH, have been chosen to refer to a therapeutic class of drugs (under *antibiotics*, under *chemical actions*), whereas *penicilline g* is considered as a chemical substance.

The structuring involved in these instances reflects the ontological commitments of the terminology designers, and cannot be recovered by lexical inclusion.[2]

4.3 'Expansion' subset

When a 'parent' term is in 'expansion' position (end position) in a 'child' term, we assume that the semantic head of the child term is modified; the induced relation is indeed expected not to be *is-a*. Some of the main cases found are close to those for the 'head' subset. Among others, we find again enumerations (see subsection 4.2.1):

immunodepresseurs / *antineoplasiques et immunodepresseurs*

and syntactic ambiguity (subsection 4.2.2):

oncogene/*antigene viral oncogene*,

where the word *oncogene* is a noun in the first term and an adjective in the second one.

Many of the relations found in the 'expansion' subset are partitive:

abdomen/*muscle droit abdomen*,
amerique centrale/*indien amerique centrale*,
argent/*nitrate argent*.

(human body parts, a continent and its peoples, and chemical substances).

In some instances, a general type of link between terms can be detected:

caused-by: *myxome*/*virus myxome*,

but in most other cases, we have what looks like a specific thematic relation between a predicate and its argument:

comportement alimentaire/*troubles comportement alimentaire*,
bovin/*pneumonie interstitielle atypique bovin*,
hopital/*capacite lits hopital*,
services sante/*fermeture service sante*,
macrophage/*activation macrophage*.

Note that some of these expansion relations involve adjectival derivations of nouns:

cubitus/*nerf cubital*,
genes/*epreuve complementation genetique*.

[1] Note, though, that if *inorganic acids* was named this way, it would be impossible to link it by lexical induction to other, more specific types of inorganic acids.

[2] They might be amenable to distributional methods if their contexts of occurrence are different enough.

4.4 'Other' subset

In this last subset, the 'parent' term can be at any position in the 'child' term other than head or expansion. It can also be non-contiguous, accepting modifiers or some other intervening elements. All these cases are actually similar to those of the 'expansion' subset except those of the form:

bacterie aerobie/*bacterie gram-negatif aerobie*

where *bacterie* remains the head of the term.

The following examples reproduce the general cases of the 'expansion' subset with additional modifiers:

arteres/*anevrysme artere iliaque*,
hepatite b/*virus hepatite b canard*,
encephalite/*virus encephalite equine ouest*,
sommeil/*troubles sommeil extrinseques*,
irrigation/*liquide irrigation endocanalaire*,
maladie/*assurance maladie personne agee*.

In some of them, adjectival derivation is involved:

cellules/*molecule-1 adhesion cellulaire vasculaire*,
chimie/*produits chimiques inorganiques*,
dent/*implantation dentaire sous-periostee*.

Some relations are characteristic of the language of chemical compounds:

cytochrome c/*ubiquinol-cytochrome c reductase*,
diphosphate/*uridine diphosphate acide glucuronique*,
lysine/*histone-lysine n-methyltransferase*.

The 'other' subset also hosted the following morphosyntactic ambiguity:

cilie/*cellule ciliee externe*

where the words *cilie* (noun, an invertebrate organism) and *ciliee* (inflected form of adjective *cilie*, which characterizes a type of cell) are conflated by lemmatization. This error is mainly due to the fact that the MeSH is written with unaccented uppercase letters: the adjective is actually spelled *cilié*, which would be unambiguous here.

5 Synthesis

We presented in this paper a human analysis of automatically, lexically-induced term relations that were not found in the terminology from which the terms were obtained (the MeSH thesaurus). This lexical method considers that a term P is probably a parent of a term C iff all the words of P occur in C. This inclusion test is helped by morphological normalization.

Morphological normalization was found to be useful not only in identifying the already existing relations (section 3.2), but also for the 'new' relations. This confirms previous work by Jacquemin and Tzoukermann (1999).

The occurrences of syntactic ambiguity suggest that morphosyntactic tagging could be useful. The methods specifically designed for detection of syntactic and morphosyntactic term variants (Bourigault, 1994; Jacquemin and Tzoukermann, 1999) might then be more efficient and less error-prone. We must be warned however that this may not be an easy task, since most of the MeSH terms are not syntactically well-formed (few determiners and prepositions, inverted heads) and contain rare, technical words that are likely to be absent from most electronic lexicons.

Spurious relations may come from several sources. A few cases are due to abusive morphological normalization; errors in term names (translation errors) were also uncovered. We made a distinction between 'head' and 'expansion' positions of the 'parent' term in its 'child'. One would expect that relations where the parent is in head position would be correct; however, this is not always true.

The putative head of a term is sometimes not correctly identified because of specific thesaural constructs (the 'comma' form) and chemical constructs (*quinone reductases* are a kind of *reductases*) which display head inversion, and because of enumerations. An additional situation is that of a term whose actual syntactic head does not entertain an *is-a* relation with it (the 'plaster cat'). Furthermore, the head word may not have a stable meaning: it may be syntactically ambiguous (*cilie*), polysemous (*investissement*) or underspecified (*acne*).

The remaining 'head' cases reveal specific modeling options, or 'ontological commitments', of the terminology designers: the relations induced might be considered semantically valid, but were discarded in the MeSH because of overall structuring choices. These choices cannot be predicted with the lexical methods used here, and seem to be the most resistant to attempts at automatic derivation. They also show that what is correct is not necessarily useful for a given terminology.

The ‘expansion’ cases may be useful to propose other relations than *is-a*: we displayed partitive relations, but left to further work a classification of the remaining ones. The UMLS semantic network relations (NLM, 2001b) might be a relevant direction to look into to represent such links.

References

Olivier Bodenreider, Anita Burgun, and Thomas C. Rindflesch. 2001. Lexically-suggested hyponymic relations among medical terms and their representation in the UMLS. In URI INIST CNRS, editor, *TIA'2001 Terminologie et Intelligence artificielle*, pages 11–21, Nancy.

Didier Bourigault. 1994. Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances à partir de textes. In *Proceedings of the 9 th Conference RFIA-AFCET*, pages 1123–1132, Paris, France, January. AFCET.

James J Cimino. 1996. Coding systems in health care. In Jan H. van Bemmel and Alexa T. McCray, editors, *Yearbook of Medical Informatics '95 — The Computer-based Patient Record*, pages 71–85. Schattauer, Stuttgart.

Natalia Grabar and Pierre Zweigenbaum. 2002. Lexically-based terminology structuring: a feasibility study. In *LREC Workshop on Using Semantics for Information Retrieval and Filtering*, pages 73–77, Las Palmas, Canaries, May. ELRA.

Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Natural Language Processing and Machine Translation. Kluwer Academic Publishers, London.

Thierry Hamon, Adeline Nazarenko, and Cécile Gros. 1998. A step towards the detection of semantic variants of terms in technical documents. In Christian Boitet, editor, *Proceedings of the 17 th COLING*, pages 498–504, Montréal, Canada, 10–14 August.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In Antonio Zampolli, editor, *Proceedings of the 14 th COLING*, pages 539–545, Nantes, France, 23–28 July.

INSERM, 2000. *Thésaurus Biomédical Français/Anglais*. Institut National de la Santé et de la Recherche Médicale, Paris.

Christian Jacquemin and Didier Bourigault. 2002. Term extraction and automatic indexing. In Ruslan Mitkov, editor, *Handbook of Computational Linguistics*. Oxford University Press, Oxford. *To appear*.

Christian Jacquemin and Évelyne Tzoukermann. 1999. NLP for term variant extraction: A synergy of morphology, lexicon, and syntax. In Tomek Strzalkowski, editor, *Natural language information retrieval*, volume 7 of *Text, speech and language technology*, chapter 2, pages 25–74. Kluwer Academic Publishers, Dordrecht & Boston.

David D. McDonald. 1993. Internal and external evidence in the identification and semantic categorization of proper names. In Branimir Boguraev and James Pustejovsky, editors, *Corpus Processing for Lexical Acquisition*, pages 61–76. MIT Press, Cambridge, MA.

Stuart J Nelson, Douglas Johnston, and Betsy L Humphreys. 2001. Relationships in medical subject headings. In Carol A Bean and Rebecca Green, editors, *Relationships in the organization of knowledge*, New York. Kluwer Academic Publishers.

National Library of Medicine, Bethesda, Maryland, 2001a. *Medical Subject Headings*. www.nlm.nih.gov/mesh/meshhome.html.

National Library of Medicine, Bethesda, Maryland, 2001b. *UMLS Knowledge Sources Manual*. www.nlm.nih.gov/research/umls/.

Patrick Séguéla and Nathalie Aussenac. 1999. Extraction de relations sémantiques entre termes et enrichissement de modèles du domaine. In Régine Teulier, editor, *Actes de IC'99*, June.

Pierre Zweigenbaum. 1999. Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé*, (2–3):27–47.