Lexically-Based Terminology Structuring: Some Inherent Limits

Natalia Grabar  Pierre Zweigenbaum

Mission de Recherche en Sciences et Technologies de l’Information Médicale (STIM)
DSI/AP–HP

{ngr,pz}@biomath.jussieu.fr
http://biomath.jussieu.fr/~pz
Outline

- Background and Objectives
- Methods
- Quantitative Results
- Qualitative Study
- Synthesis
Objective

Assess the feasibility of terminology structuring

- Start from an existing structured terminology
- Remove hierarchical links, keep set of terms
- Discover hierarchical (is-a) links
- Compare them with the original structure of the terminology
Previous work: internal methods

*Internal methods* for terminology structuring:

- **Raw word forms**  (Bodenreider et al., 2001)
- **Morphological variants**  (Jacquemin et al., 1999)
- **Syntactic structure**  (Bourigault, 1994; Jacquemin et al., 1999)
- **Semantic variants (synonyms, hyperonyms)**  (Hamon et al., 1998)
Previous work: external methods

*External methods* for terminology structuring:

- **Distributional methods** for studying similar contexts
  (Grefenstette, 1994)

- **Lexico-syntactic patterns**
  (Hearst, 1992; Séguéla and Aussenac, 1999)
This work: start with internal methods

Experiment with:

- raw word forms
- morphological variants

Method: *lexical inclusion* of terms
Lexical Inclusion Hypothesis:

- If all the words of a term are included in another term, there exists a (hierarchical) relation between these two terms.

**acides gras** | **fatty acids**

**acides gras indispensables** | **fatty acids, essential**
Experiment with the MeSH thesaurus

- Medical Subject Headings (MeSH)
  National Library of Medicine, INSERM

- Thesaurus designed for information retrieval

- 19,638 terms (“concepts”)

- Hierarchically structured
  “narrower” terms related to “broader” terms:
  26,094 child-to-parent links (direct)
  98,815 child-to-ancestor links (direct and indirect)
Term simplification

- Convert to lower case
- Remove “non-significant” words: “stop words”
- 15 (grammatical) stop words (those in MeSH) articles, prepositions:
  
  \[ au, aux, d', de, des, du, en, et, l', la, le, les, ses, un, une \]

- Remove punctuation and numbers
Morphological knowledge: inflection

Managing the various forms of a word

*intervention* – *interventions*
*abdominal* – *abdominaux*

- General lexicon (ABU)
  219,759 word pairs [lem-gen]

- Specialized medical lexicon
  2,889 word pairs [lem-med]

Lemmatization:

*intervention chirurgicale*

*interventions chirurgicales obstetricales*
Morphological knowledge: derivation

Managing derived words

aorta – aortic,
intervene – intervention

- Specialized medical lexicon
  4,517 word pairs
  (Grabar & Zweigenbaum, RIAO 2000)

aberration chromosomique, anomalies

Stemming:

aberrations chromosomes sexuels, anomalies
**Evaluation**

Compare the **results obtained** and the **original data**:

- **Links** found by lexical inclusion  
  **Terms placed** by lexical inclusion

- **Metrics**:  
  - **recall**: nb of links found (terms placed) among all correct links (terms placed)  
  - **precision**: nb of correct links (terms placed) among links found (terms placed)

- **Approaches**:  
  - **strict** (only direct parents are considered correct)  
  - **tolerant** (any ancestor is considered correct)
Summary of quantitative evaluation

See (Grabar & Zweigenbaum, LREC 2002) for details

Links

- up to 29.3% of the links found are correct (precision)
- up to 13.7% of the direct MeSH links are found by lexical inclusion (recall)

Term placement

- up to 26% of the terms were correctly placed under one of their ancestors (recall)
- the placement advices were correct in up to 58% of the cases (precision)

More morphological normalization increases recall and decreases precision as measured against existing relations in the MeSH.
A study of “spurious” relations: Outline

70% of lexically induced relations are not considered relevant by the MeSH

head
- misidentification of head
- translation error
- head is not genus
  - enumeration
  - plastic cat
- head has different senses
- ontological commitment

expansion
- enumeration
- ambiguity (POS)
- semantic relation ≠ hyperonymy
  - partitive
  - various thematic relations
other positions of head

\[ H_1 CH_2 \]
\[ C_1 H_i C_2 \]

chemical compounds

ambiguity
Head vs expansion of term

Decompose terms into head and expansion (Bourigault 1994)

absorption intestinale

Heuristic: head is leftmost word in term (in French)
Head vs expansion of term

Decompose terms into head and expansion (Bourigault 1994)

absorption intestinale
  head  expansion

Heuristic: head is leftmost word in term (in French)
Human analysis of sample “new” relations

Examine 20% sample of “new” relations between terms (total = 194)

<table>
<thead>
<tr>
<th>Normalizations</th>
<th>Head</th>
<th>Expansion</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw</td>
<td>22</td>
<td>31</td>
<td>14</td>
</tr>
<tr>
<td>lem-stem-med</td>
<td>37</td>
<td>57</td>
<td>33</td>
</tr>
</tbody>
</table>
Finding the head

Head inversion:

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

Overzealous derivational knowledge:

contracture/contraction musculaire,
biologie/testament biologique.
Term *acide linoleique alpha* is incorrect in:

```
acide linoleique/acide linoleique alpha
```

the correct spelling (and correct lexical inclusion) is:

```
acide linolenique/acide linolenique alpha
```
**Head is not genus**

For syntactic or semantic reasons

- Enumeration:

  \[\textit{acides amines / acides amines, peptides et protéines,}\]

- Change of semantic type (cf \textit{plaster cat}):

  \[\textit{personnalité/personnalité compulsive, voix/voix oesophagienne}.\]
Head is ambiguous

- Polysemy
  finance vs psychology

  *investissement*/*investissement* (psychanalyse),

chemistry vs general sense

  *absorption*/*absorption cutanee*,

default location of *goitre* is thyroid

  *goitre*/*goitre ovarien*
Head is ambiguous

- Underspecification

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

  *acides* means *inorganic acids*

  *acne* means *acne vulgaris*
Potentially correct *is-a* links purposely not included in the MeSH hierarchy:

**signs or symptoms** vs **diseases or syndroms**:

- amyotrophies/amyotrophies spinales enfance,
- hyperplasie/hyperplasie epitheliale focale,

**siblings, cousins**:

- centre public sante/centre public sante mentale,
- rectocolite/rectocolite hemorragique,
- penicillines/penicilline g.
Expansion: misidentification

- Enumeration

  immunodepresseurs / antineoplasiques et immunodepresseurs

- Syntactic ambiguity (Noun vs Adjective)

  oncogene/antigene viral oncogene
Expansion: non-is-a relations

- Partitive relations
  - *abdomen/muscle droit abdomen*,
  - *amerique centrale/indien amerique centrale*,
  - *argent/nitrate argent*

- Causal relation
  - *myxome/virus myxome*,

- Other thematic relations
  - *comportement alimentaire/troubles comportement alimentaire*,
  - *bovin/pneumonie interstitielle atypique bovin*,
  - *hopital/capacite lits hopital*,
  - *services sante/fermeture service sante*,
  - *macrophage/activation macrophage*,
  - *cubitus/nerf cubital*
Non-head non-expansion

Actually, all these cases fall into the “expansion” case except those of the form:

\textit{bacterie aerobie/bacterie gram-negatif aerobie}

where \textit{bacterie} remains the head of the term.
Non-head non-expansion \( \sim \) expansion

- \( C_1 H C_2 \) (\( C = \) child word, \( H = \) head word)

- arteres/anevrisme 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.
- cellules/molecule-1 adhesion cellulaire vasculaire,
- chimie/produts chimiques inorganiques,
- dent/implantation dentaire sous-periostee.
Chemical compounds:

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

Morphosyntactic ambiguity (aggravated by unaccented input)

- *cilie/cellule ciliée externe*

*cilie/N vs ciliée/A*
Synthesis: head relations

- Instances of excessive morphological normalization
- Errors in source terms (translation errors)
- Not all “head” relations are correct:
  - misidentification of head based on position comma form, chemical compounds, enumeration
  - “plaster cats”
  - head with variable meaning (POS ambiguity, polysemy, underspecification)
- Ontological commitments:
  specific structuring choices of terminology designers

→ Not all “head” relations are relevant


**Synthesis: general**

- Framework useful to confront actual terminology structuring usage
- Morphological knowledge helps uncover new relations (Jacquemin & Tzoukermann 1999)
- Use POS tagging?
  - would help in some cases (suppress ambiguities)
  - not an easy task here because of ungrammatical input
- A detailed analysis of spurious “head” relations
  Not all “head” relations are relevant
- “Expansion” relations:
  potential sources of other kinds of relations (Bourigault 1994)
  part-of, etc. (see also UMLS Semantic Network)
Further work

Carry on with same framework and test other methods for finding related terms:

- Synonyms
- Hypernyms
- Distributional similarity (corpus-based)