Lexical Tools
Sub-Term Mapping Tools (STMT)

Dr. Chris J. Lu

The Lexical Systems Group
NLM. LHNCBC. CGSB

April, 2012
Table of Contents

• Introduction
  ▪ MetaMap
  ▪ UMLS-Core
• STMT
  ▪ Corpus & Normalization
  ▪ Sub-Term
  ▪ Sub-Term Patterns
  ▪ Synonyms Substitutions
• Current Plan
• Questions
STMT - Introduction

Sub-term (STMT)
Sub-term

- A sub-term is a term that is a subset of another term
- A term is composed of words
- A word is delimited by space(s) or tab(s)
Sub-term: Example

- Term: otitis externa, chronic infectious
- Corpus: The SPECIALIST Lexicon
- Sub-terms:
  - otitis (E0044452)
  - otitis externa (0044453)
  - chronic (E0219527|E0016869)
  - infectious (E0044452)
Application – MetaMap*

• Longest Prefix:
  ▪ Find the longest prefix sub-term on an input term that is a lexical item (LexItem) in the SPECIALIST Lexicon
  ▪ Ignore case and punctuation

Lexical Mapping

• Algorithm:
  1. Normalization (LexItem Norm, -f:g:rs:o:l)
     - One to one
  2. Lexicon Mapping (LexAccess)
     - Table lookup (key - values)

• Example:
  University of California, Los Angeles
  ⇒ university of california los angeles (E0063261)

<table>
<thead>
<tr>
<th>Original Term from Lexicon</th>
<th>Normalized Term (key)</th>
<th>EUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>University of California, Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>University of California Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>University of California-Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Lexical Mapping (cont.)

- University of California, Los Angeles
- University of California Los Angeles
- University of California-Los Angeles
- university of california los angeles
- UNIVERSITY OF CALIFORNIA LOS ANGELES
- …

Normalization (LexItem Norm: -f:g:rs:o:l)

university of california los angeles

LexAccess (lvg –f:E)

<table>
<thead>
<tr>
<th>Original Term from Lexicon</th>
<th>Normalized Term (key)</th>
<th>EUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>University of California, Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>University of California Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>University of California-Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>University of California-Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
</tbody>
</table>
Longest Prefix

- University of California, Los Angeles is in USA …
- University of California Los Angeles is in USA …
- University of California-Los Angles is in USA …
- university of california los angeles is in USA …
- UNIVERSITY OF CALIFORNIA LOS ANGELES is in USA …
- …

Normalization (LexItem Norm: -f:g:rs:o:l)

university of california los angeles is in usa …

Lexical Mapping on each prefix sub-term
University of California, Los Angeles is in USA …

- Normalize (LexItem Norm)
  
  => university of california los angeles is in usa …

- Word by word
  
  1. university (E0063257)
  2. university of (none)
  3. university of california (E0702384)
  4. university of california, los (none)
  5. university of california, los angles (E0063261)
  6. university of california, los angles is (none)
  7. …
Longest Prefix (cont.)

- university of california los angeles is in usa …
  I. Normalized
  II. Word by word
     1. university (E0063257)
     2. university of (none)
     3. university of california (E0702384)
     4. university of california, los (none)
     5. university of california, los angles (E0063261)
     6. university of california, los angles is (none)
     7. …
- Issues:
  ▪ Performance
  ▪ Limited to 10 prefixes
Application – UMLS-CORE*

• Query expansion: to find concepts in 3 steps:
  1) Normalized lexical matches
     ▪ MRXNS_ENG.RRF: normalized string to CUI
     ▪ Lexical tools – Norm (-f:q0:g:rs:o:t:l:B:Ct:q7:q8:w)
  2) One synonymous word or phrase substitution
  3) Two synonymous word or phrase substitutions
     ▪ acronyms: AB|ANTIBODY
     ▪ British: aeroplane|airplane
     ▪ Greco-latin: cat|feline
     ▪ ECRI (devices): Airway|Tube
     ▪ Lvg: zoic|animal life
     ▪ …

UMLS-Core: Example

- OTITIS EXTERNA, CHRONIC INFECTIOUS
  1) Normalized lexical matches
     - chronic externa infectious otitis
     - chronic externon infectious otitis
     - chronic externum infectious otitis
     ⇒ No CUI found

  2) One synonymous word or phrase substitution
     ⇒ No CUI found

  3) Two synonymous word or phrase substitutions
     ⇒ …
STMT - MetaMap

• Corpus (Lexicon) – LexItem Norm

<table>
<thead>
<tr>
<th>Original Term from Lexicon</th>
<th>Normalized Term (key)</th>
<th>EUI</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>University of California, Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>University of California Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>University of California-Los Angeles</td>
<td>university of california los angeles</td>
<td>E0063261</td>
</tr>
<tr>
<td>University of California, San Francisco</td>
<td>university of california san francisco</td>
<td>E0063262</td>
</tr>
<tr>
<td>University of California SanFrancisco</td>
<td>university of california sanfrancisco</td>
<td>E0063262</td>
</tr>
<tr>
<td>University of California-SanFrancisco</td>
<td>university of california sanfrancisco</td>
<td>E0063262</td>
</tr>
<tr>
<td>University of California</td>
<td>university of california</td>
<td>E0702384</td>
</tr>
<tr>
<td>University of California, Davis</td>
<td>university of california davis</td>
<td>E0702385</td>
</tr>
<tr>
<td>University of California Davis</td>
<td>university of california davis</td>
<td>E0702385</td>
</tr>
<tr>
<td>University of California-Davis</td>
<td>university of california davis</td>
<td>E0702385</td>
</tr>
<tr>
<td>University of California, San Diego</td>
<td>university of california san diego</td>
<td>E0702386</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

• Load the corpus to a tree structure with each term as a branch in the tree and each word in the term as a node in the branch
STMT

Corpus (Normalized) → Sub-Terms → Sub-Term Patterns → Synonyms Substitutions

$Root

university

of

california

los

angles

$End
Corpus (Normalized) -> Sub-Terms -> Sub-Term Patterns -> Synonyms Substitutions

$Root

university
of
california
los
angles
francisco
san

$End $End
Corpus (Normalized) → Sub-Terms → Sub-Term Patterns → Synonyms Substitutions

$Root

university

of

california

los

san

angles

francisco

diego

$End

$End

$End
STMT – Longest Prefix

Corpus (Normalized) → Sub-Terms → Sub-Term Patterns → Synonyms Substitutions

university of california los angeles is in usa …
university of california los angeles is in usa ...
university of california los angeles is in usa …

STMT – Longest Prefix

Corpus (Normalized) → Sub-Terms → Sub-Term Patterns → Synonyms Substitutions

$Root

california

of

university

los

angles

illinois

california

san

sanfrancisco

los

francisco

diego

angles

$End

$End

$End
university of california los angeles is in usa …
university of california los angeles is in usa …
university of california los angeles is in usa …
university of california los angeles is in usa …
university of california los angeles is in usa ...

STMT – Sub-terms

Corpus (Normalized) ➔ Sub-Terms ➔ Sub-Term Patterns ➔ Synonyms Substitutions

university of california los angeles is in usa ...

$Root

california

of

university

los

angles

illinois

california

los

san

sanfrancisco

angles

francisco

diego

$End

$End

$End

$End

$End

$End
OTITIS EXTERNA, CHRONIC INFECTIOUS

1) Normalized lexical matches
   • chronic externa infectious otitis => no CUI found
   • chronic externon infectious otitis => no CUI found
   • chronic externum infectious otitis => no CUI found
2) One sub-term (synonymous word or phrase) substitution
3) Two sub-term substitutions
STMT – UMLS-Core

- Corpus (Synonyms files):
  - Acronyms, British, Greco-latin, ECRI, Lvg, etc.
- Synonym Norm (-f: g:rs:Ct:o:l)
  - Abstract away from punctuation, cases, spelling variants, inflectional variants
  - One to many
  - Remove from synonym list if the synonym pair are the same after norm (side-to-side | side to side)
- Load the normalized synonyms to corpus tree
<table>
<thead>
<tr>
<th>Original Term from Synonyms</th>
<th>Normalized Term (key)</th>
<th>synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>chronic</td>
<td>chronic</td>
<td>chron</td>
</tr>
<tr>
<td>chronic</td>
<td>chronic</td>
<td>long-term</td>
</tr>
<tr>
<td>chronic</td>
<td>chronic</td>
<td>persistent</td>
</tr>
<tr>
<td>chronic</td>
<td>chronic</td>
<td>recurrent periodic</td>
</tr>
<tr>
<td>chronic</td>
<td>chronic</td>
<td>relapsing</td>
</tr>
<tr>
<td>infectious</td>
<td>infectious</td>
<td>communicable</td>
</tr>
<tr>
<td>infectious</td>
<td>infectious</td>
<td>contagious</td>
</tr>
<tr>
<td>infectious</td>
<td>infectious</td>
<td>infection</td>
</tr>
<tr>
<td>otitis externa</td>
<td>otitis externa</td>
<td>auditory</td>
</tr>
<tr>
<td>otitis externa</td>
<td>otitis externa</td>
<td>auditory canal</td>
</tr>
<tr>
<td>otitis externa</td>
<td>otitis externa</td>
<td>aural</td>
</tr>
<tr>
<td>otitis externa</td>
<td>otitis externa</td>
<td>ear</td>
</tr>
<tr>
<td>ANTIBODY</td>
<td>antibody</td>
<td>AB</td>
</tr>
<tr>
<td>Airway</td>
<td>airway</td>
<td>Tube</td>
</tr>
<tr>
<td>non-profit</td>
<td>non profit</td>
<td>not for profit</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
STMT – UMLS-Core

Corpus (Normalized) → Sub-Terms → Sub-Term Patterns → Synonyms Substitutions

$Root

- airway
- antibody
- chronic
- infectious
- non
- otitis
- profit
- externa
- ...

$End

$End

$End

$End

$End

$End

$End
• OTITIS EXTERNA, CHRONIC INFECTIOUS
  => otitis externa chronic infectious
• Three sub-terms found
  ▪ otitis externa|0|2
  ▪ chronic|2|3
  ▪ infectious|3|4
• OTITIS EXTERNA, CHRONIC INFECTIOUS
  • => otitis externa chronic infectious

• Sub-term patterns with one substitution
  ▪ **otitis externa** chronic infectious
  ▪ otitis externa **chronic** infectious
  ▪ otitis externa chronic **infectious**

• Sub-term patterns with two substitutions
  ▪ **otitis externa chronic** infectious
  ▪ **otitis externa** chronic **infectious**
  ▪ otitis externa **chronic infectious**

• It could be really complicated (multiple overlaps)
• otitis externa chronic infectious

• Synonyms:
  ▪ otitis externa (4): auditory | auditory canal | aural | ear
  ▪ chronic (5): chron | long-term | persistent | relapsing | recurrent periodic
  ▪ infectious (3): communicable | contagious | infection

• Sub-term patterns with one substitution (12)
  ▪ **otitis externa** chronic infectious (4)
  ▪ otitis externa **chronic** infectious (5)
  ▪ otitis externa chronic **infectious** (3)

• Sub-term patterns with two substitutions (47)
  ▪ **otitis externa chronic** infectious (20)
  ▪ **otitis externa chronic infectious** (12)
  ▪ otitis externa **chronic infectious** (15)
• otitis externa chronic infectious

• Sub-term patterns with one substitution (12)
  ▪ **otitis externa** chronic infectious (4)
  ▪ otitis externa **chronic** infectious (5)
  ▪ otitis externa chronic **infectious** (3)
  => No CUI Found

• Sub-term patterns with two substitutions (47)
  ▪ **otitis externa chronic** infectious (20)
  ▪ **otitis externa chronic infectious** (12)
  ▪ otitis externa **chronic infectious** (15)
  => **ear** chronic **infection** | C0743359 | EAR INFECTION CHRONIC
STMT Plan

• First release: STMT, 2013
• Provide Java APIs, command line tools
  ▪ Generic tool
    o STMT (Sub-Term Mapping Tool)
  ▪ Preloaded corpus
    o LSF (LexItem Sub-Term Finder)
    o SMT (Synonym Mapping Tool)
    o …
• Web site with full documentations
Current Status/Future Work

- Major software components are completed
- Future work:
  - Tests
  - Web site - documents
  - Deployable package
  - Persistent corpus
References


Questions