NLM Indexing Initiative Tools for NLP: MetaMap and the Medical Text Indexer

Natural Language Processing: State of the Art, Future Directions

April 23, 2012

Alan R. Aronson
Outline

• Introduction
• MetaMap
  • Overview
  • Linguistic roots
  • Recent Word Sense Disambiguation (WSD) efforts
• The NLM Medical Text Indexer (MTI)
  • Overview
  • MTI as First-line Indexer (MTIFL)
  • Recent improvements
  • Gene indexing
MetaMap/MTI Example

- **MetaMap** identifies biomedical concepts in text

> Cigarette smoking increases the mean platelet volume in elderly patients with risk factors for atherosclerosis.

- **Medical Text Indexer (MTI)** summarizes text using MetaMap and the Medical Subject Headings (MeSH) vocabulary

- Cigarette Smoking
- Tobacco
- Blood Platelets
- Aged
- Humans
- Risk Factors
- Arteriosclerosis
- Atherosclerosis
Outline

• Introduction
• MetaMap
  • Overview
  • Linguistic roots
  • Recent Word Sense Disambiguation (WSD) efforts
• The NLM Medical Text Indexer (MTI)
  • Overview
  • MTI as First-line Indexer (MTIFL)
  • Recent improvements
  • Gene indexing
MetaMap Overview

• Named-entity recognition program
• Identify UMLS Metathesaurus concepts in text
• Linguistic rigor
• Flexible partial matching
• Emphasis on thoroughness rather than speed
The MetaMap Algorithm

- Parsing
  - Using SPECIALIST minimal commitment parser, SPECIALIST lexicon, MedPost part of speech tagger

- Variant generation
  - Using SPECIALIST lexicon, Lexical Variant Generation (LVG)

- Candidate retrieval
  - From the Metathesaurus

- Candidate evaluation

- Mapping construction
MetaMap Evaluation Function

- Weighted average of
  - centrality (is the head involved?)
  - variation (average of all variation)
  - coverage (how much of the text is matched?)
  - cohesiveness (in how many pieces?)
### MetaMap Processing Example

**Inferior vena caval stent filter** (PMID 3490760)

**Candidate Concepts:**

- **C0080306:** Inferior Vena Cava Filter
- **C0180860:** Filter
- **C0581406:** Filter
- **C1522664:** Filter
- **C1704449:** Filter (function)
- **C1704684:** Filter Device Component
- **C1875155:** FILTER
- **C0521360:** Inferior vena caval
- **C0042460:** Vena caval
- **C0038257:** Stent
- **C1705817:** Stent Device Component
- **C0447122:** Vena

**Metathesaurus UMLS Semantic Type**

<table>
<thead>
<tr>
<th>UMLS Semantic Type</th>
<th>MetaMap Score</th>
<th>Metathesaurus String</th>
<th>UMLS Semantic Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>[medd]</td>
<td>C0080306</td>
<td>Inferior Vena Cava Filter</td>
<td>[mnob]</td>
</tr>
<tr>
<td>[medd]</td>
<td>C0180860</td>
<td>Filter</td>
<td>[mnob]</td>
</tr>
<tr>
<td>[medd]</td>
<td>C0581406</td>
<td>Filter</td>
<td>[medd]</td>
</tr>
<tr>
<td>[medd]</td>
<td>C1522664</td>
<td>Filter</td>
<td>[inpr]</td>
</tr>
<tr>
<td>[cnce]</td>
<td>C1704449</td>
<td>Filter (function)</td>
<td>[cnce]</td>
</tr>
<tr>
<td>[medd]</td>
<td>C1704684</td>
<td>Filter Device Component</td>
<td>[medd]</td>
</tr>
<tr>
<td>[medd]</td>
<td>C0038257</td>
<td>Stent</td>
<td>[medd]</td>
</tr>
<tr>
<td>[medd]</td>
<td>C1705817</td>
<td>Stent Device Component</td>
<td>[medd]</td>
</tr>
<tr>
<td>[bpoc]</td>
<td>C0447122</td>
<td>Vena</td>
<td>[bpoc]</td>
</tr>
</tbody>
</table>
MetaMap Final Mappings

Inferior vena caval stent filter

Final Mappings (subsets of candidate sets):

Meta Mapping (911):
909  C0080306: Inferior Vena Cava Filter [medd]
637  C1705817: Stent                     [medd]

Meta Mapping (911):
909  C0080306: Inferior Vena Cava Filter [medd]
637  C0038257: Stent                     [medd]
Word Sense Disambiguation (WSD)

- Kids with *colds* may also have a sore throat, cough, headache, mild fever, fatigue, muscle aches, and loss of appetite.

- Candidate MetaMap mappings for *cold*
  
  - C0234192: Cold (Cold sensation)
  - C0009264: Cold (Cold temperature)
  - C0009443: Cold (Common cold)
Knowledge-based WSD

- Compare **UMLS** candidate concept profile vectors to context of ambiguous word
- Concept profile vectors’ words from definition, synonyms and related concepts

<table>
<thead>
<tr>
<th>Common cold</th>
<th>Cold temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight</td>
<td>Word</td>
</tr>
<tr>
<td>265</td>
<td>infect</td>
</tr>
<tr>
<td>126</td>
<td>disease</td>
</tr>
<tr>
<td>41</td>
<td>fever</td>
</tr>
<tr>
<td>40</td>
<td>cough</td>
</tr>
<tr>
<td>258</td>
<td>temperature</td>
</tr>
<tr>
<td>86</td>
<td>hypothermia</td>
</tr>
<tr>
<td>72</td>
<td>effect</td>
</tr>
<tr>
<td>48</td>
<td>hot</td>
</tr>
</tbody>
</table>

- Candidate concept with highest similarity is predicted
Knowledge-based WSD

• Kids with *colds* may also have a sore throat, *cough*, headache, mild *fever*, fatigue, muscle aches, and loss of appetite.

<table>
<thead>
<tr>
<th>Common cold</th>
<th>Weight</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>265</td>
<td>infect</td>
</tr>
<tr>
<td></td>
<td>126</td>
<td>disease</td>
</tr>
<tr>
<td></td>
<td>41</td>
<td><em>fever</em></td>
</tr>
<tr>
<td></td>
<td>40</td>
<td><em>cough</em></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cold temperature</th>
<th>Weight</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>258</td>
<td><em>temperature</em></td>
</tr>
<tr>
<td></td>
<td>86</td>
<td><em>hypothermia</em></td>
</tr>
<tr>
<td></td>
<td>72</td>
<td><em>effect</em></td>
</tr>
<tr>
<td></td>
<td>48</td>
<td><em>hot</em></td>
</tr>
</tbody>
</table>
Automatically Extracted Corpus WSD

- **MEDLINE** contains numerous examples of ambiguous words context, though not disambiguated

Candidate concept: common cold

Unambiguous synonyms:
- "common cold"[tiab]
- "acute nasopharyngitis"[tiab] …

Query:
- CUI:C0009443
- CUI:C0009264
- "cold temperature"[tiab]
- "low temperature"[tiab] …

PubMed
WSD Method Results

- Corpus method has better accuracy than UMLS method
  - MSH WSD data set created using MeSH indexing
    - 203 ambiguous words
    - 81 semantic types
    - 37,888 ambiguity cases
- Indirect evaluation with summarization and MTI correlates with direct evaluation

<table>
<thead>
<tr>
<th></th>
<th>UMLS</th>
<th>Corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>NLM WSD</td>
<td>0.65</td>
<td>0.69</td>
</tr>
<tr>
<td>MSH WSD</td>
<td>0.81</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Outline

• Introduction
• MetaMap
  • Overview
  • Linguistic roots
  • Recent Word Sense Disambiguation (WSD) efforts
• The NLM Medical Text Indexer (MTI)
  • Overview
  • MTI as First-line Indexer (MTIFL)
  • Recent improvements
  • Gene indexing
Cigarette smoking increases factors for atherosclerosis.

Kario K, Matsuo T, Nakao K.
Department of Internal Medicine, Hyogo Prefectural

Abstract
To study the effects of cigarette smoking and (MPV) and other platelet parameters in 142
The MPV and the platelet count were higher
0.54, P < 0.05) when compared with the non-
in 8 smoking subjects in the atherosclerotic
suggest that smoking may increase platelet
megakaryocytes are activated to produce large
smoking may also contribute to the acceleration
atherosclerotic disease.
MTI

- MetaMap Indexing – Actually found in text
- Restrict to MeSH – Maps UMLS Concepts to MeSH
- PubMed Related Citations – Not necessarily found in text

Related Citations
Title + Abstract
Ordered list of MeSH Main Headings
MeSH Main Headings
UMLS concepts
Clustering & Ranking
Restrict to MeSH
PubMed
Related Citations
Extract MeSH Descriptors
Final Ordered list of MeSH Headings
Apply Indexing Rules
CheckTag Expansion
Subheading Attachment

Received 2,330 Indexer Feedbacks
Incorporated 40% into MTI

Hibernation *should only be indexed for animals, not for "stem cell hibernation"

Clove *(spice) should not be mapped to the verb "cleave"
MTI Uses

• Assisted indexing of MEDLINE by Index Section
• Assisted indexing of Cataloging and History of Medicine Division records
• Automatic indexing of NLM Gateway meeting abstracts
• First-line indexing (MTIFL) since February 2011
MTI as First-Line Indexer (MTIFL)

“Normal” MTI Processing

MTI Processes/Recommends MeSH

Indexer Reviews Selects

Reviser Reviews Selects Adjusts Approves

Indexing Displays in PubMed as Usual
MTI as First-Line Indexer (MTIFL)

MTI Processes/Indexes MeSH

Indexer Reviews Selects

Reviser Reviews Selects Adjusts Approves

Indexing Displays in PubMed as Usual

Index Section Compares MTI and Reviser Indexing

MTIFL MTI Processing

45 MEDLINE Journals
CheckTags Machine Learning Results

- 200k citations for training and 100k citations for testing

<table>
<thead>
<tr>
<th>CheckTag</th>
<th>F₁ before ML</th>
<th>F₁ with ML</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle Aged</td>
<td>1.01%</td>
<td>59.50%</td>
<td>+58.49</td>
</tr>
<tr>
<td>Aged</td>
<td>11.72%</td>
<td>54.67%</td>
<td>+42.95</td>
</tr>
<tr>
<td>Child, Preschool</td>
<td>6.11%</td>
<td>45.40%</td>
<td>+39.29</td>
</tr>
<tr>
<td>Adult</td>
<td>19.49%</td>
<td>56.84%</td>
<td>+37.35</td>
</tr>
<tr>
<td>Male</td>
<td>38.47%</td>
<td>71.14%</td>
<td>+32.67</td>
</tr>
<tr>
<td>Aged, 80 and over</td>
<td>1.50%</td>
<td>30.89%</td>
<td>+29.39</td>
</tr>
<tr>
<td>Young Adult</td>
<td>2.83%</td>
<td>31.63%</td>
<td>+28.80</td>
</tr>
<tr>
<td>Female</td>
<td>46.06%</td>
<td>73.84%</td>
<td>+27.78</td>
</tr>
<tr>
<td>Adolescent</td>
<td>24.75%</td>
<td>42.36%</td>
<td>+17.61</td>
</tr>
<tr>
<td>Humans</td>
<td>79.98%</td>
<td>91.33%</td>
<td>+11.35</td>
</tr>
<tr>
<td>Infant</td>
<td>34.39%</td>
<td>44.69%</td>
<td>+10.30</td>
</tr>
<tr>
<td>Swine</td>
<td>71.04%</td>
<td>74.75%</td>
<td>+3.71</td>
</tr>
</tbody>
</table>
CheckTags Machine Learning Results

- 200k citations for training and 100k citations for testing

<table>
<thead>
<tr>
<th>CheckTag</th>
<th>F_1 before ML</th>
<th>F_1 with ML</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle Aged</td>
<td>1.01%</td>
<td>59.50%</td>
<td>+58.49</td>
</tr>
<tr>
<td>Aged</td>
<td>11.72%</td>
<td>54.67%</td>
<td>+42.95</td>
</tr>
<tr>
<td>Child, Preschool</td>
<td>6.11%</td>
<td>45.40%</td>
<td>+39.29</td>
</tr>
<tr>
<td>Adult</td>
<td>19.49%</td>
<td>56.84%</td>
<td>+37.35</td>
</tr>
<tr>
<td>Male</td>
<td>38.47%</td>
<td>71.14%</td>
<td>+32.67</td>
</tr>
<tr>
<td>Aged, 80 and over</td>
<td>1.50%</td>
<td>30.89%</td>
<td>+29.39</td>
</tr>
<tr>
<td>Young Adult</td>
<td>2.83%</td>
<td>31.63%</td>
<td>+28.80</td>
</tr>
<tr>
<td>Female</td>
<td>46.06%</td>
<td>73.84%</td>
<td>+27.78</td>
</tr>
<tr>
<td>Adolescent</td>
<td>24.75%</td>
<td>42.36%</td>
<td>+17.61</td>
</tr>
<tr>
<td>Humans</td>
<td>79.98%</td>
<td>91.33%</td>
<td>+11.35</td>
</tr>
<tr>
<td>Infant</td>
<td>34.39%</td>
<td>44.69%</td>
<td>+10.30</td>
</tr>
<tr>
<td>Swine</td>
<td>71.04%</td>
<td>74.75%</td>
<td>+3.71</td>
</tr>
</tbody>
</table>
## CheckTags Machine Learning Results

- 200k citations for training and 100k citations for testing

<table>
<thead>
<tr>
<th>CheckTag</th>
<th>$F_1$ before ML</th>
<th>$F_1$ with ML</th>
<th>Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Middle Aged</td>
<td>1.01%</td>
<td>59.50%</td>
<td>+58.49</td>
</tr>
<tr>
<td>Aged</td>
<td>11.72%</td>
<td>54.67%</td>
<td>+42.95</td>
</tr>
<tr>
<td>Child, Preschool</td>
<td>6.11%</td>
<td>45.40%</td>
<td>+39.29</td>
</tr>
<tr>
<td>Adult</td>
<td>19.49%</td>
<td>56.84%</td>
<td>+37.35</td>
</tr>
<tr>
<td>Male</td>
<td>38.47%</td>
<td>71.14%</td>
<td>+32.67</td>
</tr>
<tr>
<td>Aged, 80 and over</td>
<td>1.50%</td>
<td>30.89%</td>
<td>+29.39</td>
</tr>
<tr>
<td>Young Adult</td>
<td>2.83%</td>
<td>31.63%</td>
<td>+28.80</td>
</tr>
<tr>
<td>Female</td>
<td>46.06%</td>
<td>73.84%</td>
<td>+27.78</td>
</tr>
<tr>
<td>Adolescent</td>
<td>24.75%</td>
<td>42.36%</td>
<td>+17.61</td>
</tr>
<tr>
<td>Humans</td>
<td>79.98%</td>
<td>91.33%</td>
<td>+11.35</td>
</tr>
<tr>
<td>Infant</td>
<td>34.39%</td>
<td>44.69%</td>
<td>+10.30</td>
</tr>
<tr>
<td>Swine</td>
<td>71.04%</td>
<td>74.75%</td>
<td>+3.71</td>
</tr>
</tbody>
</table>
MTI - How are we doing?

Focus on Precision versus Recall

Fruition of 2011 Changes
**FLNA filamin A, alpha [Homo sapiens]**

Gene ID: 2316, updated on 10-Mar-2012

**Summary**

- **Official Symbol**: FLNA provided by HGNC
- **Official Full Name**: filamin A, alpha provided by HGNC
- **Primary source**: HGNC:3754
- **Locus tag**: XX-FW83128A1.1
- **See related**: Ensembl:ENSG00000196924, HPRD:02060, MIM:300017, Vega:OTTHUMG00000022712
- **Gene type**: protein coding
- **RefSeq status**: REVIEWED
- **Organism**: Homo sapiens
- **Lineage**: Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorhini; Catarrhini; Hominidae; Homo
- **Also known as**: FLN, FMD, MNS, OPD, ABPX, CVD1, FLN1, NHBP, OPD1, OPD2, XLVD, XMVD, FLN-A, ABP-280

The protein encoded by this gene is an actin-binding protein that crosslinks actin filaments and links actin filaments to membrane glycoproteins. The encoded protein is involved in remodeling the cytoskeleton to effect changes in cell shape and migration. This protein interacts with integrins, transmembrane receptor complexes, and second messengers. Defects in this gene are a cause of several syndromes, including periventricular nodular heterotopias (PVNH1, PVNH4).

**GeneRIFs: Gene References Into Functions**

1. These results demonstrate that FLNA is prone to pathogenic rearrangements
2. mutations in FLNA may represent an unrecognized cause of macrothrombocytopenia with an altered platelet production and a modified platelet-vessel wall interaction
3. study reports on two brothers with X-linked cardiac valvular dystrophy and a hemizygous FLNA mutation and review previously described cases from the literature
4. Consistent with structural predictions, strain increases beta-integrin binding to FLNA, whereas it causes FliGAP to dissociate from FLNA, providing a direct and specific molecular basis for cellular mechanotransduction
5. Hepatitis C virus nonstructural (NS) 3 and NS5A proteins were associated with filamin A, while core protein partially with filamin A and vimentin
6. regulates actin-linked caveolae dynamics following loss of cell adhesion
7. Adapter protein SH2B1beta binds filamin A to regulate prolactin-dependent cytoskeletal reorganization and cell motility
8. crystal structure of FlnA Ig10 determined at 2.44 A resolution provides insight into the perturbations caused by these mutations
9. The presence of these clinical findings in a mutation-confirmed case of OPD2 supports the notion that corneal clouding, bifid tongue, and DWM are part of the constellation of
The Gene Indexing Assistant (GIA)

- An automated tool to assist the indexer in identifying and creating GeneRIFs
  - Evaluate the article
  - Identify genes
  - Make links to Entrez Gene
  - Suggest geneRIF annotation

- Anticipated Benefits:
  - Increase in speed
  - Increase in comprehensiveness
The NLM Indexing Initiative Team

- Alan R. Aronson (Project Leader)
- James G. Mork (Staff)
- François-Michel Lang (Staff)
- Willie J. Rogers (Staff)
- Antonio J. Jimeno-Yepes (Postdoctoral Fellow)
- J. Caitlin Sticco (Library Associate Fellow)