Retrieving Medical Records with SENNAMED: NEC Labs America at TREC 2012 Medical Record Track

Yanjun Qi and Pierre-Francois Laquerre

Department of Machine Learning, NEC Laboratories America Princeton, New Jersey, USA

yanjun@nec-labs.com, pierre.francois@nec-labs.com
NECLA “sennaMed” System

- Three text analytic modules (we believe) are necessary for “Semantic Retrieval of Medical Records”

- (I) Medical report understanding
  - Medical concept extraction and categorization
    - “knee surgery” is a procedure
    - “nausea” is a symptom
  - Report information extraction
    - Extract implicit information, e.g. the patient’s gender even if “male” or “female” are not explicitly mentioned
    - Tell the difference between the patient’s history, her family history, her current state

- (II) Query text understanding
  - Similar to “report understanding” as above

- (III) Classic IR retrieval ➔ query vs. medical records
sennaMed (I): Report understanding

We need to interpret the reports into meaningful medical profiles

Term Transformation to UMLS Semantic Concepts

A 46-year-old lady was referred to the haematology clinic for evaluation of lymphocytosis in May 1993. She had severe lethargy and intermittent right upper abdominal discomfort with significant loss in weight. Her past medical history included essential hypertension controlled on atenolol 100 mg once daily and was also on frusemide 40 mg once daily. She had no significant surgical history other than having undergone cholecystectomy in 1972. She is not a smoker. Physical examination showed no evidence of lymphadenopathy. Peripheral blood showed numerous Howell-Jolly bodies within erythrocytes. Thyroid function tests, protein electrophoresis, C-reactive protein, immunoglobulin levels and autoimmune screening were normal. Ultrasonography and computed tomography scan of the abdomen and pelvis did not reveal mediastinal lymphadenopathy, but the spleen was noted to be very atrophic. Gastroscopy showed multiple gastric erosions and the initial impression was of peptic disease complicated by lymphoma and lymphocytosis.
sennaMed (II): Query understanding

Similarly, we need to interpret the query topic

1) Women with hearing loss
2) Patients admitted for care who take herbal products for osteoarthritis
3) Diabetic patients who received diabetic education in the hospital
4) Cancer patients with liver metastasis treated in the hospital who underwent a procedure
5) Patients with GERD who had esophageal adenocarcinoma diagnosed by endoscopy
sennaMed (III): Retrieval Modules

• Two classic retrieval models for indexing and ranking
  – (1) a vector space retrieval model and
  – (2) a language model based retrieval approach.

• Also test several other classic IR techniques, which include
  – (1) dimension reduction using latent semantic indexing, and
  – (2) query expansions.
Our four submissions

<table>
<thead>
<tr>
<th>Official Runs</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>sennamed1</td>
<td>UMLS concept representation, language model retrieval, query expansion</td>
</tr>
<tr>
<td>sennamed2</td>
<td>UMLS concept representation, vector space retrieval, query expansion</td>
</tr>
<tr>
<td>sennamed3</td>
<td>UMLS concept representation, vector space retrieval</td>
</tr>
<tr>
<td>sennamedlsi</td>
<td>UMLS concept based representation, vector space retrieval, LSI</td>
</tr>
</tbody>
</table>

Parameters are tuned on the 2011 TREMed test data.
Also tried other methods (SSI, LDA, query expansion variations, etc.)
Metamap: medical concepts extracted as words, with negated concepts included as separate terms (e.g. “smoker” is different than “not a smoker”)
## Global results

<table>
<thead>
<tr>
<th>Metric</th>
<th>sennamed1</th>
<th>sennamed2</th>
<th>sennamed3</th>
<th>sennamed1LSI</th>
</tr>
</thead>
<tbody>
<tr>
<td>infAP</td>
<td>0.2246</td>
<td><strong>0.2745</strong></td>
<td>0.2169</td>
<td>0.2151</td>
</tr>
<tr>
<td>infNDCG</td>
<td>0.478</td>
<td><strong>0.5468</strong></td>
<td>0.4688</td>
<td>0.4468</td>
</tr>
<tr>
<td>P@10</td>
<td>0.5255</td>
<td><strong>0.5574</strong></td>
<td>0.5447</td>
<td>0.4468</td>
</tr>
<tr>
<td>R-prec</td>
<td>0.3457</td>
<td><strong>0.3805</strong></td>
<td>0.3298</td>
<td>0.2974</td>
</tr>
<tr>
<td>bpref</td>
<td>0.3647</td>
<td><strong>0.427</strong></td>
<td>0.3559</td>
<td>0.3496</td>
</tr>
<tr>
<td>recip_rank</td>
<td>0.7706</td>
<td><strong>0.7696</strong></td>
<td>0.7497</td>
<td>0.6189</td>
</tr>
</tbody>
</table>
sennaMed2 run (3\textsuperscript{rd} overall / 2\textsuperscript{nd} in automatic runs)

<table>
<thead>
<tr>
<th># topics</th>
<th>infAP</th>
<th>infNDCG</th>
<th>R-prec</th>
<th>P@10</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best</strong></td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>11</td>
</tr>
<tr>
<td>Better than median</td>
<td>27</td>
<td>27</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td><strong>Median</strong></td>
<td>2</td>
<td>1</td>
<td>6</td>
<td>12</td>
</tr>
<tr>
<td>Below median</td>
<td>13</td>
<td>13</td>
<td>10</td>
<td>7</td>
</tr>
<tr>
<td><strong>Worst</strong></td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td><strong>Summary (good/bad)</strong></td>
<td>34 / 13</td>
<td>34 / 13</td>
<td>35 / 12</td>
<td>36 / 11</td>
</tr>
</tbody>
</table>
EXTRA about Concept Detection Using MetaMap

Courtesy slides from
MetaMap 2010 Tutorial
Why Concept Identification?

- Structured data vs. text
- Concept identification is useful/essential for many tasks including
  - Information extraction/Data mining
  - Classification/Categorization
  - Text summarization
  - Question answering
  - Literature-based knowledge discovery

Courtesy slides from MetaMap 2010 Tutorial
Example (best mappings with WSD)

- PMID – 9339686
- AB – Cerebral blood flow (CBF) in newborn infants is often below levels necessary to sustain brain viability in adults.

Courtesy slides from MetaMap 2010 Tutorial
The 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

Courtesy slides from MetaMap 2010 Tutorial
NLM Applications using MetaMap: RIDeM (Repository for Informed Decision Making)

- Summarization
- Meta-analysis, Reviews
- Drugs (indications, interactions, etc.)
- Clinical question answering (CQA 1.0)
- Information for patients
- Linking evidence to patient records (InfoBot)
- Image retrieval (iMEDLINE)
- Annotation of Interactive Publications
- Clinical research (HDiscovery)
- Translational research: linking of basic research to clinical information

Courtesy slides from MetaMap 2010 Tutorial