Cohort Shepherd: Discovering Cohort Traits from Hospital Visits

Travis Goodwin, Bryan Rink, and Sanda Harabagiu
Presented by: Sanda Harabagiu

The University of Texas at Dallas
Human Language Technology Research Institute
http://www.hlt.utdallas.edu
Outline

- **Our Approach**+ The system architecture
- Keyword extraction
- Distillation of Topic requirements
- Query expansion
- Relevant visit retrieval
- Evaluation and analysis
- Conclusion
Content-based retrieval for EMRs

- **Given:**
  1. EMRs from Pitts’ BLULAB NLP Repository
  2. Mappings between hospital visits and EMRs
  3. Topics that target certain cohorts

- **Expected:** ranked list of hospital visits

- **Resources:** SnowMed, UMLS, PubMed Central

- **The approach:**
  1. Converting topics into queries
  2. Query expansion
  3. Relevant Visit Retrieval
  4. Result filtering
The architecture of Cohort Shepherd

- Topics
  - Keyword Extraction
    - Age Recognition
    - Gender Recognition
    - Negation Extraction
      - Disjunction Detection
        - Query Expansion
          - Structured Query
            - Retrieved
              - Lucene Index
              - Indri Index
                - Ranked List of Visits
                  - Filtering
                    - Re-Ranked List of Visits

- Medical Records
- Report to Visit Mapping
- Visit Documents

- WIKIPEDIA
- PubMed
- SNOMED
- UMLS
Keyword extraction

- Identify medical concepts or conditions in each topic
  - Noun chunks extracted using phrase chunkers (parsers) available from the OpenNLP toolkit
  - Identify title of Wikipedia articles within each noun chunk
  - E.g. "lower extremity chronic wound"
    - "lower extremity" and "chronic wound" are both Wikipedia articles
    - Extract the phrases as two keyword sequences
  - Fallback extraction method if no Wikipedia article titles are seen in the noun phrase
Fallback for Keyword extraction

- Extract the longest disjoint phrases from a noun chunk which do not:
  - Occur more than 300,000 times in PubMed Central Open Access Subset
  - Start with \{a, any, her, the, some, to, etc. \}
  - Contain a non-content word (e.g., visit, patient, doctor, appointment, record)
  - Consist of a single letter or only punctuation

- **Example:** “patients admitted with new-onset diabetes”
  - diabetes – from Wikipedia
  - new-onset – from the fallback method
Outline

- Our Approach: The system architecture
- Keyword extraction
- **Distillation of Topic requirements**
- Query expansion
- Relevant visit retrieval
- Evaluation and analysis
- Conclusion
Distillation of Topic Requirements

- Identification of patient age requirements
  - “elderly patients with ventilator-associated pneumonia”
  - “patients in their 20s and 30s admitted for overdose”
- Extracted using a set of regular expression
- Checked retrieved records for “**AGE[in 50s]”, “**AGE[in 60s]”, “**AGE[in 70s]”, etc. to match de-identified ages in the records
Distillation of Topic Requirements

- Extraction of **patient gender** requirements
  - “*men with prostate cancer treated with surgery*”
  - Regular expressions
    - 17 different gender indicators
    - E.g., *man, men, boy, he*, etc.
  - Detect gender in *both the topic and retrieved visits*
    - Discard documents containing the wrong gender more frequently than the gender mandated by the topic
Outline

- Our Approach: The system architecture
- Keyword extraction
- Distillation of Topic requirements
- **Query expansion**
- Relevant visit retrieval
- Evaluation and analysis
- Conclusion
Query expansion

- **UMLS**
  - Expand keyword to any other concepts with the same UMLS concept ID (synonyms)
  - Examples:
    - *stroke* → apoplexy, brain attack, “vascular accident, brain”, cerebrovascular accident, accident – cerebrovascular
    - *lower extremity* → hindlimb, hind limb, lower leg, leg, leg region

- **SNOMED CT**
  - Expand along *is_a, part_of*, and *component* relations
  - Examples:
    - *atypical antipsychotics* → abilify, aripiprazole, asenapine, clozapine, clozaril
    - *dementia* → alzheimer’s disease, vascular dementia, dialysis dementia, neurosyphilis, postconcussion syndrome
Query expansion

- Wikipedia redirect expansion
  - Expand to synonyms, alternate or mis-spellings of keywords
  - Examples:
    - hearing loss ➔ auditory impairment, deaf, deafness, hard of hearing, hearing damage
    - ablation ➔ ablate, ablated, ablative cooling, ablative material, rotoablation
Query expansion

- Co-occurrence expansion
  - Expand keyword with related terms
  - Examples:
    - lower extremity → ankle, amputation, tibial, thigh, toe, popliteal (hollow at the back of knee)
    - atypical antipsychotics → olanzapine, risperidone, quetiapine, clozapine, antipsychotic drug
  - Co-occurrence strength is calculated using normalized Google distance (Glibris and Vitanyi, 2007):
    \[
    NGD(x, y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}
    \]
    - Where \( M \) is the number of documents, \( x \) and \( y \) are terms in the document, \( f(x) \) and \( f(y) \) are the document frequencies of terms \( x \) and \( y \), and \( f(x, y) \) is the document frequency of the terms co-occurring
Outline

- Our Approach+ The system architecture
- Keyword extraction
- Distillation of Topic requirements
- Query expansion
  - Relevant visit retrieval
- Evaluation and analysis
- Conclusion
Relevant visit retrieval

- Four submitted runs
  - **Judged**
    - Standard Lucene scoring
    - Multiple keyword focused scoring:
      - \[ \text{score}(q, d) = \sum_{k \in R} \text{Lucene}(\text{query}(k), d) \times \prod_{k \in R} \text{Lucene}(\text{query}(k), d) + |R| \]
      - Where \( q \) is a query, \( d \) is a document, \( k \) is a keyword, and \( R \) is the set of keywords occurring in both \( q \) and \( d \).
  - **Non-judged**
    - Hybrid approach (Lucene + Indri)
      - With NegEx for negation
    - Hybrid with LingScope for negation
Outline

- Our Approach
- The system architecture
- Keyword extraction
- Distillation of Topic requirements
- Query expansion
- Relevant visit retrieval
- Evaluation and analysis
- Conclusion
## Evaluation and analysis

<table>
<thead>
<tr>
<th>Description</th>
<th>Run Tag</th>
<th>P10</th>
<th>MAP</th>
<th>BPref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Lucene scoring</td>
<td>UTDHLTSL</td>
<td>56.2%</td>
<td>35.6%</td>
<td>50.8%</td>
</tr>
<tr>
<td>Multiple keyword focused scoring</td>
<td>UTDHLTMK</td>
<td>63.2%</td>
<td>40.2%</td>
<td>54.3%</td>
</tr>
<tr>
<td>Hybrid Approach with NegEx</td>
<td>UTDHLTCIR</td>
<td>60.3%</td>
<td>40.8%</td>
<td>54.5%</td>
</tr>
<tr>
<td>Hybrid Approach with LingScope</td>
<td>UTDHLTCIRLS</td>
<td>60.0%</td>
<td>40.0%</td>
<td>53.4%</td>
</tr>
</tbody>
</table>

- **Where:**
  - P10 is the percent of relevant documents in the first 10 results
  - MAP is the mean average precision
  - BPref is the binary preference score
## Experiments with the Hybrid Approach:

<table>
<thead>
<tr>
<th>Description</th>
<th>MAP</th>
<th>Diff.</th>
<th>BPref</th>
<th>Diff.</th>
</tr>
</thead>
<tbody>
<tr>
<td>With Negation Filtering using Negex</td>
<td>0.4082</td>
<td>0.0000</td>
<td>0.5449</td>
<td>0.0000</td>
</tr>
<tr>
<td>With Negation Filtering using LingScope</td>
<td>0.3995</td>
<td>-0.0087</td>
<td>0.5339</td>
<td>-0.0110</td>
</tr>
<tr>
<td>Without Negation Filtering</td>
<td>0.4165</td>
<td>0.0083</td>
<td>0.5481</td>
<td>0.0032</td>
</tr>
<tr>
<td>Without Age Filtering</td>
<td>0.4107</td>
<td>0.0025</td>
<td>0.5479</td>
<td>0.0030</td>
</tr>
<tr>
<td>Without Gender Filtering</td>
<td>0.4052</td>
<td>-0.0030</td>
<td>0.5439</td>
<td>-0.0010</td>
</tr>
<tr>
<td>Without Co-occurrence Expansion</td>
<td>0.3540</td>
<td>-0.0542</td>
<td>0.4934</td>
<td>-0.0515</td>
</tr>
<tr>
<td>Without SNOMED expansion</td>
<td>0.4085</td>
<td>0.0003</td>
<td>0.5497</td>
<td>0.0048</td>
</tr>
<tr>
<td>Without UMLS expansion</td>
<td>0.4034</td>
<td>-0.0048</td>
<td>0.5435</td>
<td>-0.0014</td>
</tr>
<tr>
<td>Without Wikipedia expansion</td>
<td>0.4036</td>
<td>-0.0046</td>
<td>0.5451</td>
<td>0.0002</td>
</tr>
</tbody>
</table>
Outline

- Our Approach + The system architecture
- Keyword extraction
- Distillation of Topic requirements
- Query expansion
- Relevant visit retrieval
- Evaluation and analysis
- Conclusion
Outline

- Keyword extraction
- Topic requirement extraction
- Query expansion
- Relevant visit retrieval
- Evaluation and analysis
- Conclusion
Conclusion

1. Query expansion helped !!!
   - Co-occurrence information provided large gain
   - Ability to determine related keywords, not just synonymous terms

2. LingScope negation and speculation information could be used to filter negative results with better parameterization

3. Difficulty in predicting at training time the nature of test topics
   - Spent time solving problems which did not help for test data (e.g., age and gender cues)

4. Availability of relevance assessments will allow for more focused efforts in the future
Questions?

Thank You!