CONCEPT-BASED AND MULTIMODAL METHODS FOR MEDICAL CASE RETRIEVAL

SUPPLEMENT
PhD Defense
Mario Taschwer
May 17, 2017
1. Introduction
   - Medical Case Retrieval (MCR)
   - Problem Statement
   - Contributions
2. Processing Compound Figures
3. Biomedical Concept Mapping
4. Using Concepts for textual MCR
5. Multimodal MCR
6. Further Work
Medical Case Retrieval (MCR)

- Major component of medical decision support systems based on case-based reasoning
- Solution may help to generate datasets for medical education and research
**Problem Statement**

- **State of the art for MCR on general datasets:**
  - Best systems employ purely textual techniques
- **Main research problem:**
  - How to improve MCR methods using textual and visual information?
- **Hypothesis:**
  - Biomedical concepts may help – with techniques:
    - Query or document expansion for text retrieval
    - Concept-based retrieval
    - Fusion of text and concept-based retrieval
CONTRIBUTIONS OF PhD THESIS

• Novel automatic methods for compound figure classification and separation
• Evaluation of concept mapping techniques:
  ▪ New and existing methods of mapping text or images to biomedical concepts
• Comparison of query and document expansion by biomedical concepts for textual MCR
• Novel framework combining text and concept-based retrieval, improving over state of the art
1. Introduction

2. Processing Compound Figures
   - Classification
   - Separation
   - Combined evaluation

3. Biomedical Concept Mapping

4. Using Concepts for textual MCR

5. Multimodal MCR

6. Further Work
COMPOUND FIGURES

Subfigures of article images are separated by:

- edges
- or
- whitespace

- Compound figure classification (CFC)
- Automatic separation (CFS)
- Chained CFC and CFS
**Compound Figure Classifier**

Is a given image a compound figure?

- **Proposed features:** spatial profiles of projections
  - Projected values: intensity statistics, Hough transform
- **Machine learning:** logistic regression, SVM
- **Evaluation** on ~10,000 images: 76.9% accuracy
  - Inferior to state of the art (82.8%)
  - But more efficient: 12.3 images per second (MATLAB)
**Compound Figure Separation**

- Accuracy on ~3400 images: 84.9%
  - better than best known semi-automatic result (84.6%)
CFC-CFS Chain

- Chain accuracy on ~6800 images:
  - Without CFC: 85.1%
  - With "best" CFC: 87.3% (low false negative rate)
  - With ideal CFC: 92.5%
OUTLINE 3

1. Introduction
2. Processing Compound Figures
3. Biomedical Concept Mapping
   - Medical Subject Headings (MeSH)
   - Text-to-Concept Mapping
   - Image-to-Concept Mapping
4. Using Concepts for textual MCR
5. Multimodal MCR
6. Further Work
**Medical Subject Headings**

- Thesaurus of biomedical concepts:
  - ~27k primary terms, ~161k synonyms
  - “More general than” relations between primary terms impose 16 tree structures (maximal depth 11)
- Used to index biomedical publications
  - MeSH annotations created by domain experts

<table>
<thead>
<tr>
<th>Primary MeSH Term</th>
<th>Node Identifier</th>
<th>Specialty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abortion, Spontaneous</td>
<td>C13.703.039</td>
<td>2</td>
</tr>
<tr>
<td>Pregnancy Complications</td>
<td>C13.703</td>
<td>1</td>
</tr>
<tr>
<td>Female Urogenital Diseases and Pregnancy Complications</td>
<td>C13</td>
<td>0</td>
</tr>
</tbody>
</table>
TEXT-TO-CONCEPT MAPPING 1

• Existing systems:
  - MetaMap, Open Biomedical Annotator: slow
  - Whatizit MeshUp: kNN classifier, short text input only

• Novel, more efficient string matching approach:
  - based on inverted index of MeSH terms
  - finds (partial) occurrences of MeSH terms in single pass through text document

• Effectiveness evaluated for two objectives:
  - classification: reproducing manual MeSH annotations
  - concept-based retrieval on MCR dataset (~75k docs)
Text-to-Concept Mapping 2

Text classification of 1000 documents (title, abstract)

Efficiency:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MetaMap</td>
<td>4,30</td>
</tr>
<tr>
<td>BinDist</td>
<td>0,01</td>
</tr>
<tr>
<td>OBA</td>
<td>11,8</td>
</tr>
<tr>
<td>MeshUp</td>
<td>N/A</td>
</tr>
</tbody>
</table>

String matching

- MetaMap
- BinDist
- OBA

kNN classifier

- MeshUp

Graph showing efficiency of various algorithms.
**TEXT-TO-CONCEPT MAPPING 3**

Concept-based retrieval on MCR dataset

MGT: “ideal” concept mapping using ground-truth MeSH terms

35 queries
75k docs
BinDist index

3 Algorithms used for query mapping
**Image-to-Concept Mapping**

**M1:** visual kNN

**M2:** concept-based kNN + visual reranking

**M3:** concept-based kNN (no visual information)

Collects MeSH terms from image index (figure captions, CEDD features, 300k images)

Concept-based retrieval on MCR dataset (35 queries, 75k docs BinDist index)
OUTLINE 4

1. Introduction
2. Processing Compound Figures
3. Biomedical Concept Mapping
4. Using Concepts for textual MCR
   ▪ Query Expansion
   ▪ Document Expansion
5. Multimodal MCR
6. Further Work
Query / Document Expansion 1

• Query expansion:
  ▪ Expand textual query with additional relevant terms:
    • MeSH terms resulting from concept mapping
    • Discriminative terms from pseudo-relevant documents (pseudo-relevance feedback)
  ▪ Perform text retrieval with expanded query

• Document expansion:
  ▪ Expand full text of documents with relevant MeSH terms prior to indexing
Text retrieval on MCR dataset

<table>
<thead>
<tr>
<th>Method</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>F+</td>
<td>0.16</td>
</tr>
<tr>
<td>F</td>
<td>0.17</td>
</tr>
<tr>
<td>M</td>
<td>0.18</td>
</tr>
<tr>
<td>M+</td>
<td>0.19</td>
</tr>
<tr>
<td>Fr</td>
<td>0.20</td>
</tr>
<tr>
<td>Fr+</td>
<td>0.21</td>
</tr>
<tr>
<td>Mr</td>
<td>0.22</td>
</tr>
<tr>
<td>B13</td>
<td>0.23</td>
</tr>
<tr>
<td>Mr+</td>
<td>0.24</td>
</tr>
</tbody>
</table>

- F: full text
- M: MeSH query expansion (BinDist)
- B13: best result at ImageCLEF 2013
- MeSH document expansion
- Pseudo-relevance feedback
1. Introduction
2. Processing Compound Figures
3. Biomedical Concept Mapping
4. Using Concepts for textual MCR
5. Multimodal MCR
   - Framework for Text- and Concept-Based Retrieval
   - Fusion Methods
   - Results
6. Further Work
RETRIEVAL FRAMEWORK

- **indexing**
  - build text index
  - text index
  - medical text concept classifier
  - medical image concept classifier
  - Biomedical concepts
  - MCR dataset
  - build concept index
  - concept index

- **retrieval**
  - case query
  - query text
  - query images
  - detect concepts
  - text retrieval
  - expanded query
  - detect concepts
  - query concepts
  - textual query expansion
  - concept selection
  - query expansion
  - pseudo-relevance feedback
  - ranked document list
  - rank/score fusion
  - result list
  - concept-based retrieval
FUSION METHODS

• Fuse result lists of retrieval methods A and B

• Linear fusion: \[ s = \beta \cdot s_A + (1 - \beta) \cdot s_B \]
  ▪ Combine retrieval scores with fixed weight \( \beta \)
  ▪ \( s_A, s_B \): logistic score normalization from rank positions

• Query-adaptive fusion (QAF):
  ▪ For each query \( q \), choose weight \( \beta \) depending on \( q \)
  ▪ E.g. by estimating performance of A and B for \( q \)
    \[ \beta = \frac{p_A^2}{p_A^2 + p_B^2} \]
  ▪ “Ideal” QAF: use an oracle returning true average precision for \( p_A \) and \( p_B \)
LINEAR FUSION

T: text retrieval with query and document expansion (weight $\beta$)
C: concept-based retrieval (textual kNN concept mapping)
C+: concept-based retrieval with ground-truth MeSH terms
FUSION RESULTS

L: linear fusion with optimized weight
Q: ideal query-adaptive fusion
F: fulltext retrieval
OUTLINE 6

1. Introduction
2. Processing Compound Figures
3. Biomedical Concept Mapping
4. Using Concepts for textual MCR
5. Multimodal MCR
6. Limitations and Further Work
Limitations of MCR Dataset 1

Retrieved judged documents per query

- C
- T
- L(T,C)
- Q(T,C)
- C+
- L(T,C+)
- Q(T,C+)
LIMITATIONS OF MCR DATASET 2

Distribution of relevant judged documents per query
LIMITATIONS OF MCR DATASET 3

- Ground-truth MeSH annotations:
  - Only 77% of documents (~57k) are annotated
  - MeSH annotations tend to be incomplete and biased by domain of expertise of human annotators
  - No MeSH annotations of images in MCR dataset

- Additional relevance judgments and MeSH annotations are needed for future work
FURTHER WORK 1

• Image preprocessing:
  - Classification and filtering of diagnostic images
  - Classify modalities of diagnostic images:
    e.g. ultrasound, MRI, CT, X-ray
  - Classification of body parts represented in diagnostic images (IRMA code)
  - Apply deep learning techniques to these problems
Further Work 2

• Concept mapping:
  - Extended evaluation of string matching and image-to-concept mapping algorithms
  - Utilize other biomedical vocabularies and ontologies
  - Evaluate concept mapping by multi-view learning
  - Perform a study of manual MeSH annotations
  - Acquire an MCR dataset with more complete ground-truth MeSH annotations and relevance judgments
  - Apply deep learning to concept mapping (recent advances in image caption generation)
FURTHER WORK 3

- Text-based retrieval:
  - Utilize document structure
    (title, abstract, image captions)
  - Apply more sophisticated query expansion methods
    - Use external text corpora
    - Apply text categorization methods based on machine learning
FURTHER WORK 4

• Practical query-adaptive fusion:
  ▪ Estimate query performance of component systems from their ranking scores
  ▪ Consider other performance weighting schemes or fusion strategies

• Retrieval in multi-view latent space:
  ▪ Latent space created by subspace learning techniques may be used for direct retrieval
  ▪ Assumption: nearby points in latent space represent semantically similar cases
Further Work 5

• Learning from medical expert users:
  ▪ Use relevance feedback for short-term or long-term learning
  ▪ Apply transductive (semi-supervised) techniques for long-term learning, e.g. manifold-ranking
  ▪ Consider active learning approaches to cope with the small sample size problem for long-term learning
Publications 1


