Concept-Based and Multimodal Methods for Medical Case Retrieval

SUPPLEMENT
PhD Defense

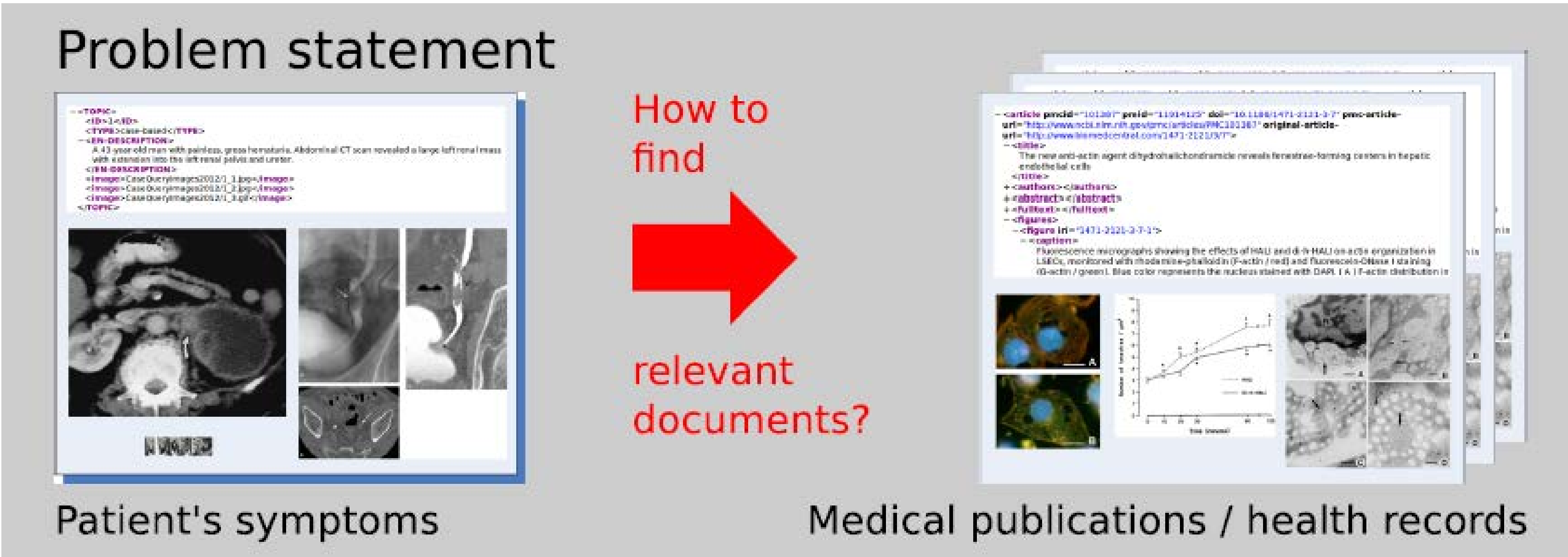
Mario Taschwer

May 17, 2017

Outline 1

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

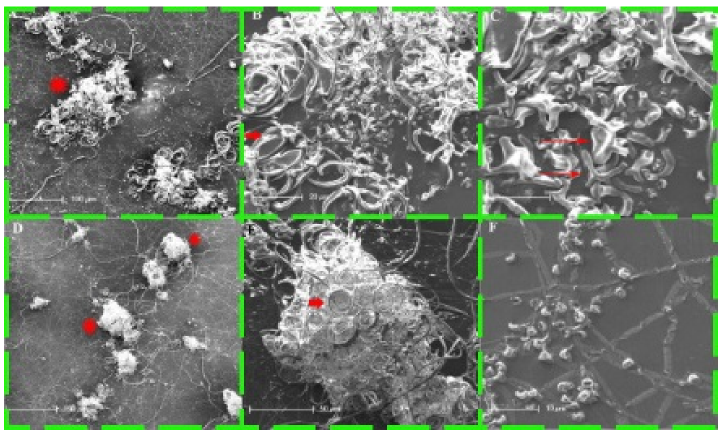
- Novel automatic methods for compound figure classification and separation
- Evaluation of concept mapping techniques:
  - New and existing methods of mapping text or images to biomedical concepts
- Comparison of query and document expansion by biomedical concepts for textual MCR
- Novel framework combining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

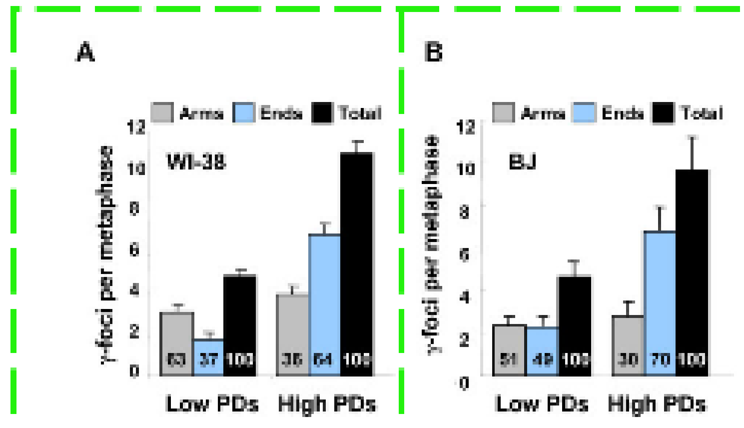
Outline 2

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

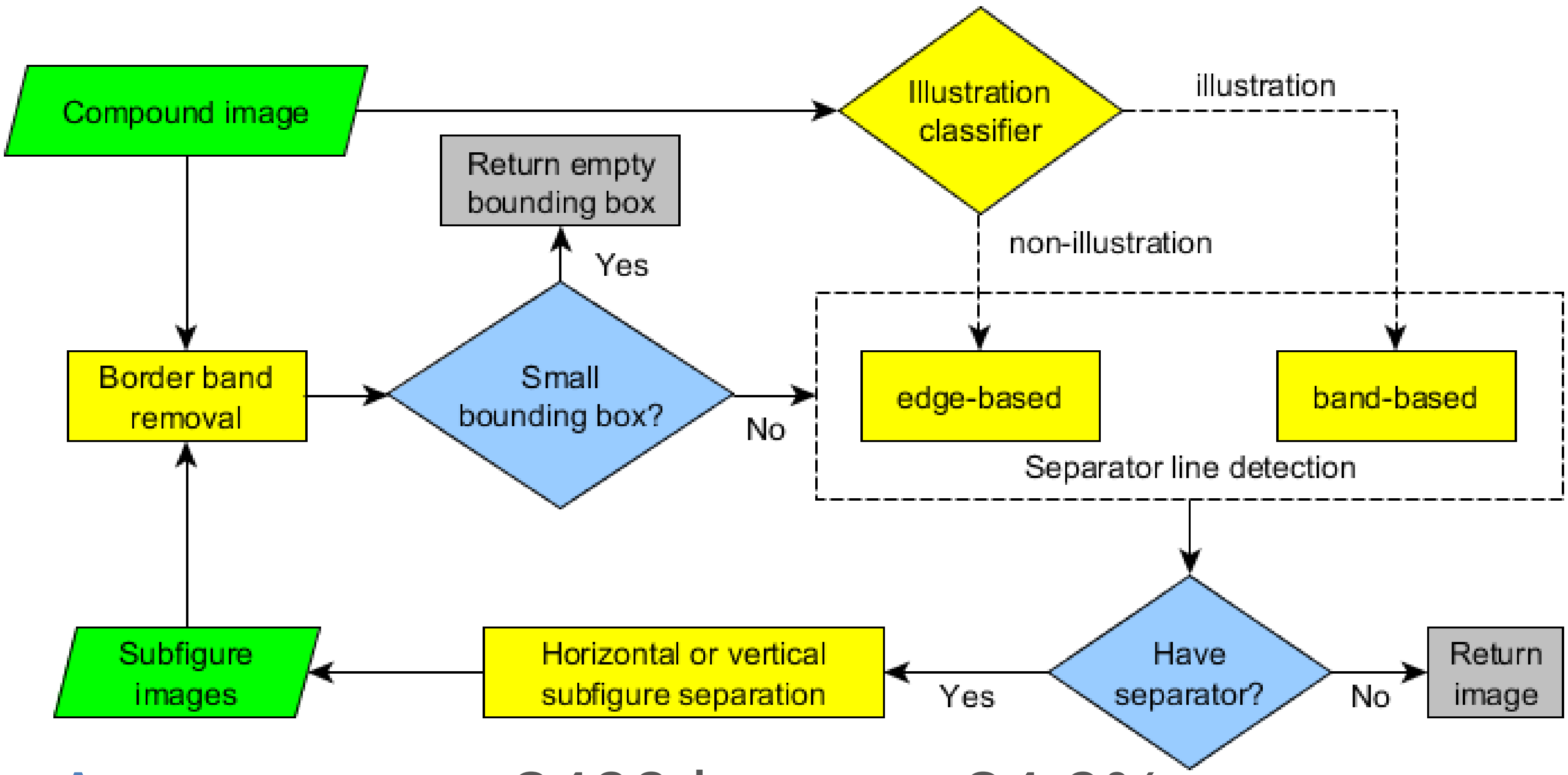
- Compound figure classification (CFC)
- Automatic separation (CFS)
- Chained CFC and CFS

Compound Figure Classifier

Is a given image a compound figure?

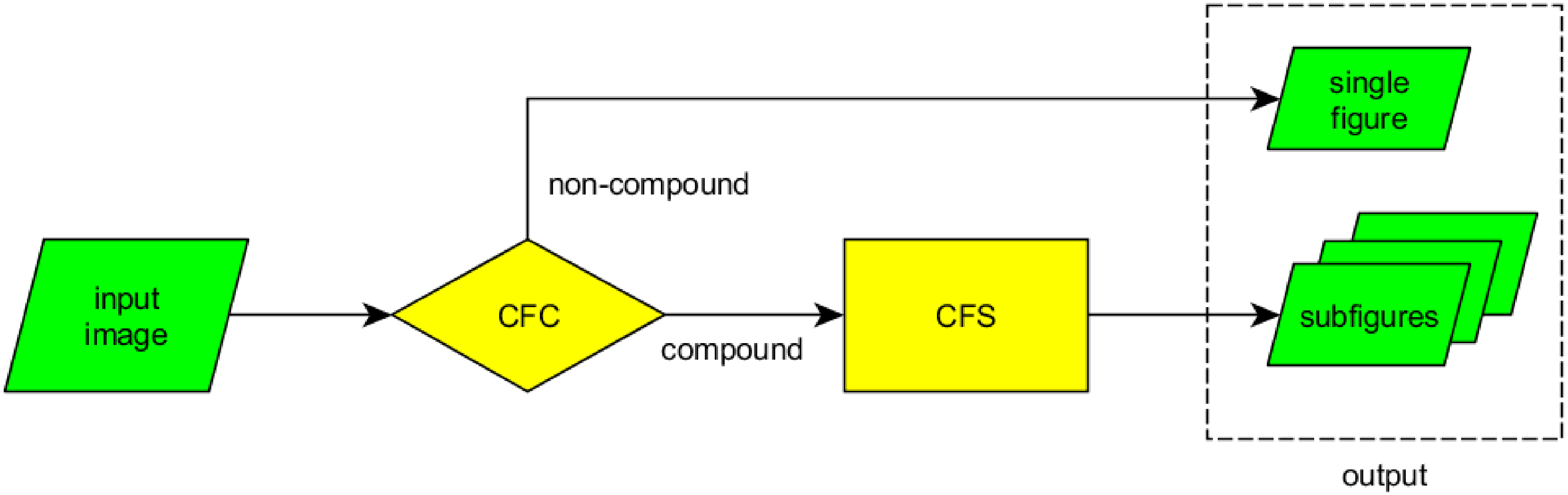
- Proposed features: spatial profiles of projections
  - Projected values: intensity statistics, Hough transform
- Machine learning: logistic regression, SVM
- Evaluation on ~10,000 images: 76.9% accuracy
  - inferior to state of the art (82.8%)
  - but more efficient: 12.3 images per second (MATLAB)

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

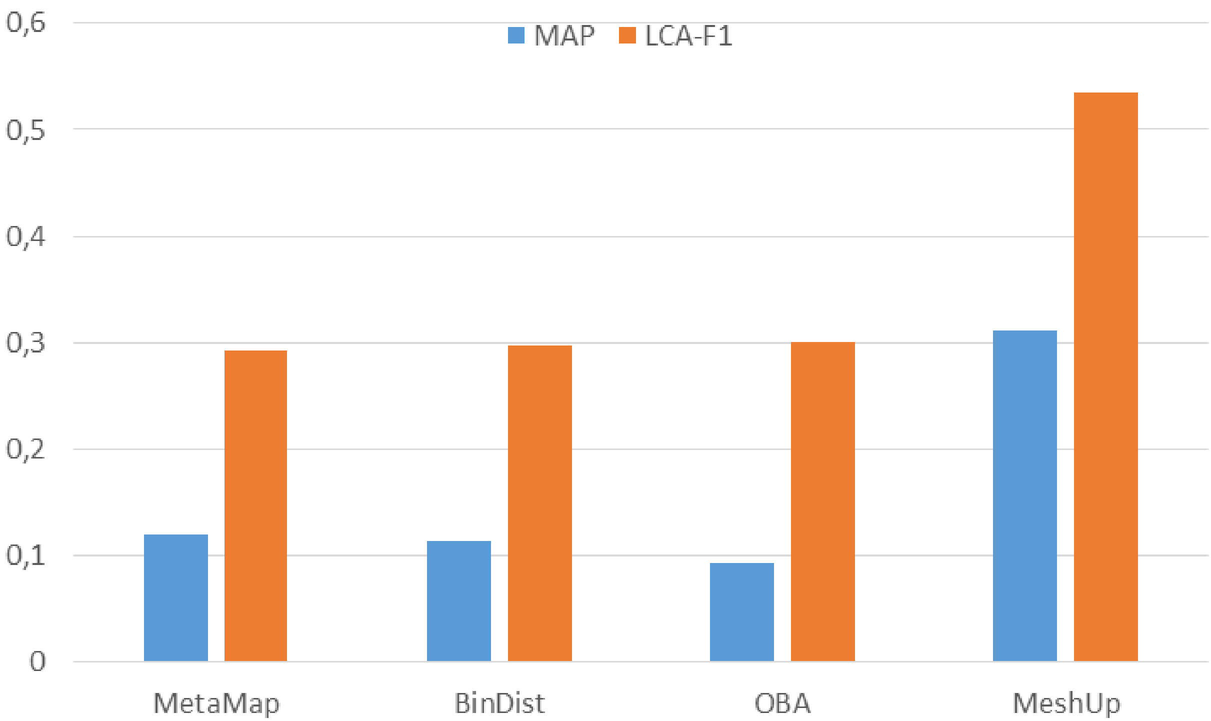
Primary MeSH Term	Node Identifier	Specialty
Abortion, Spontaneous	C13.703.039	2
Pregnancy Complications	C13.703	1
Female Urogenital Diseases and Pregnancy Complications	C13	0

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)



String matching kNN classifier

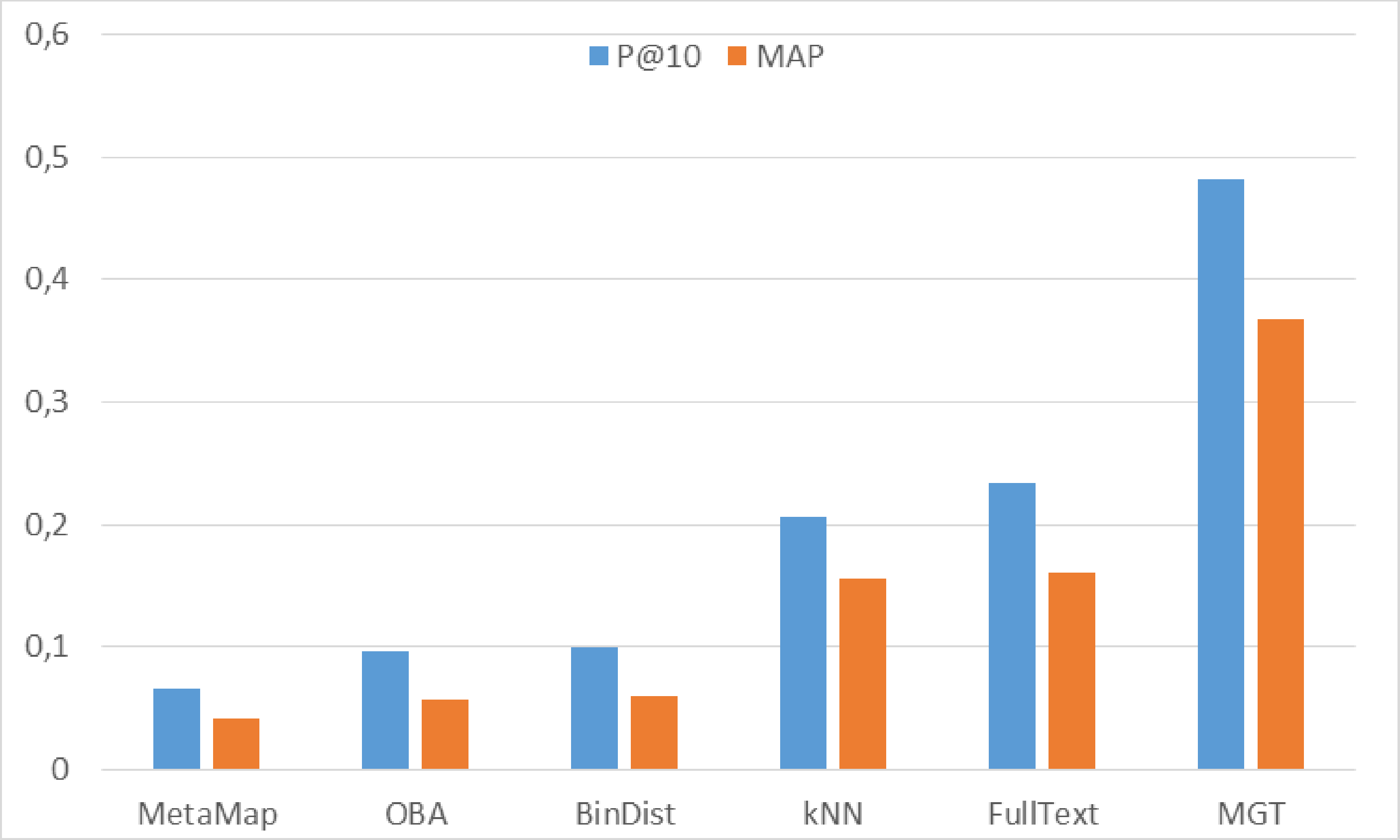
Efficiency:

Algorithm	Time (s)
MetaMap	4,30
BinDist	0,01
OBA	11,8
MeshUp	N/A

TEXT-TO-CONCEPT MAPPING 3

Concept-based retrieval on MCR dataset

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



35 queries

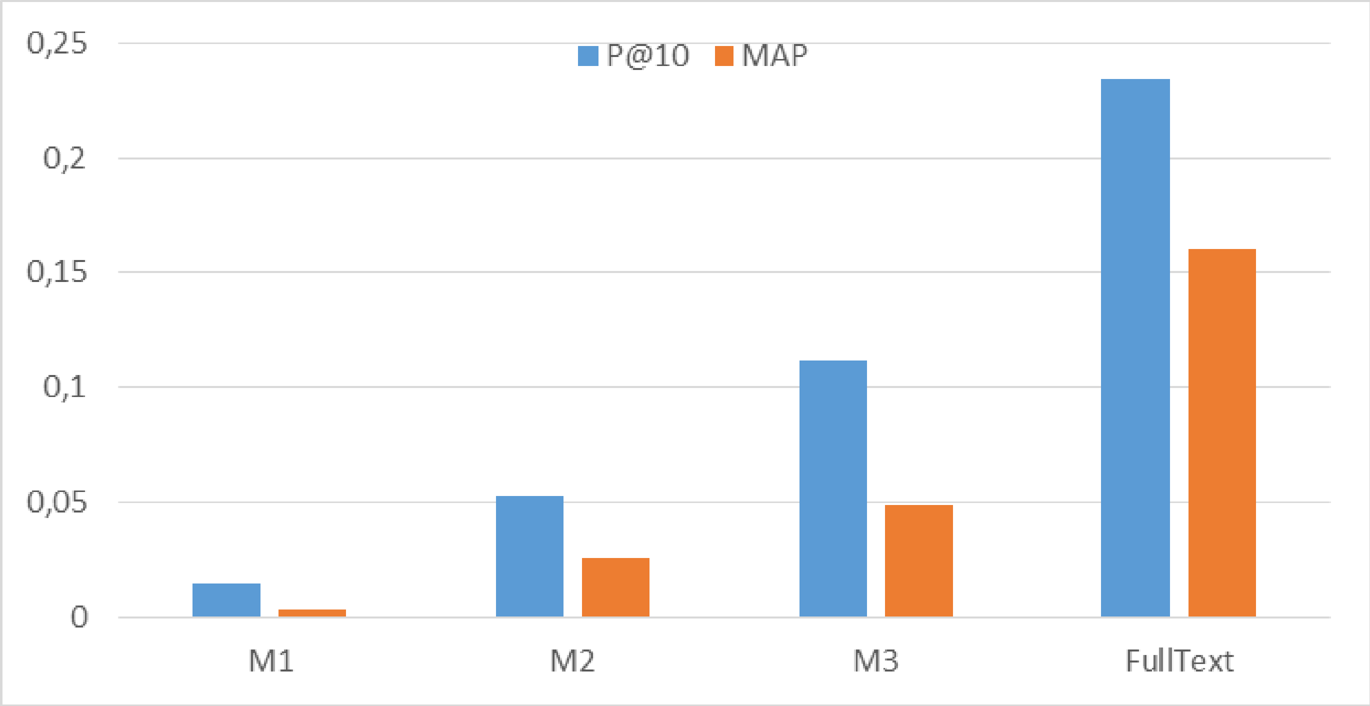
75k docs

BinDist index

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)

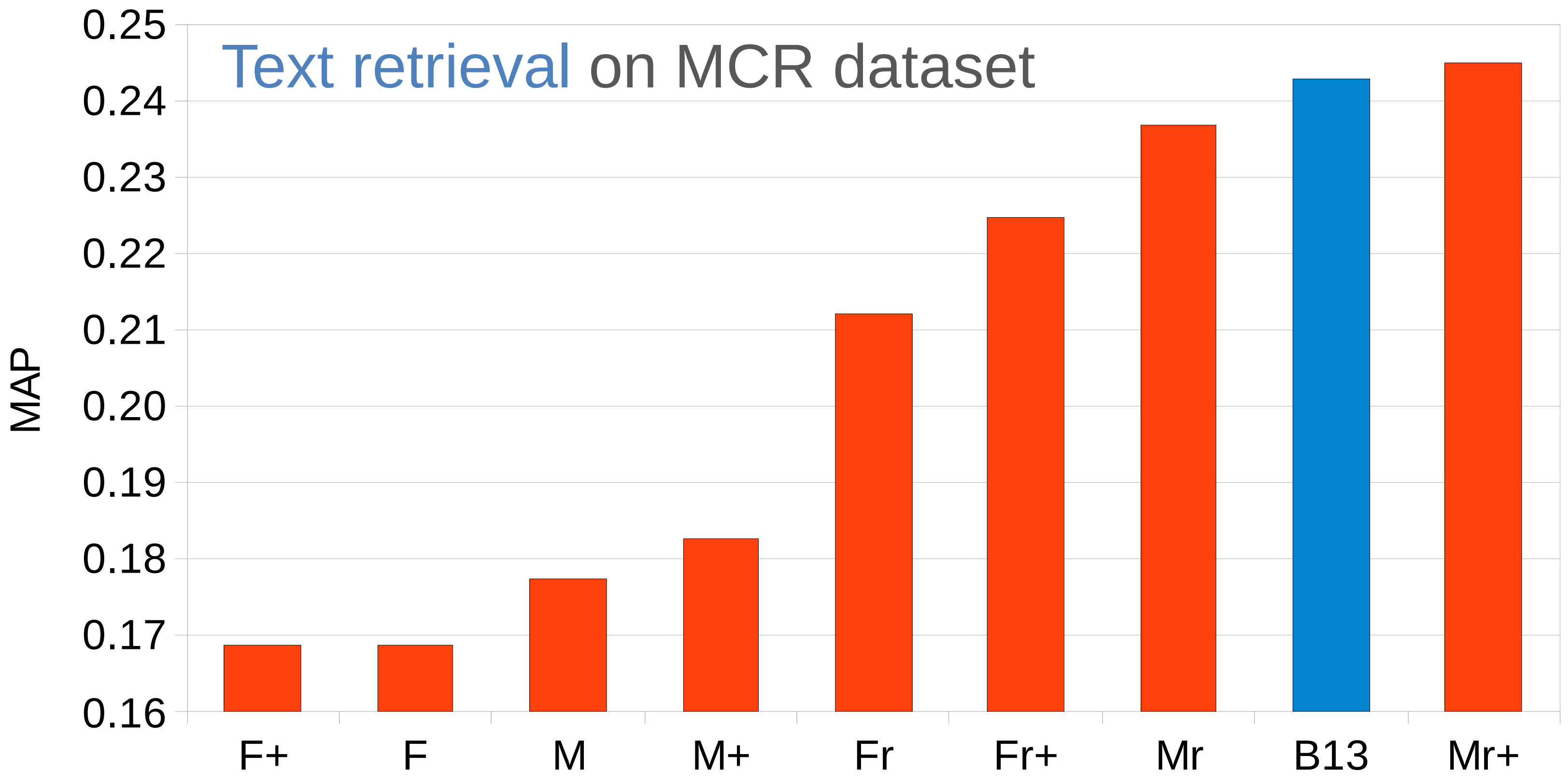
OUTLINE 4

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

QUERY / DOCUMENT EXPANSION 2

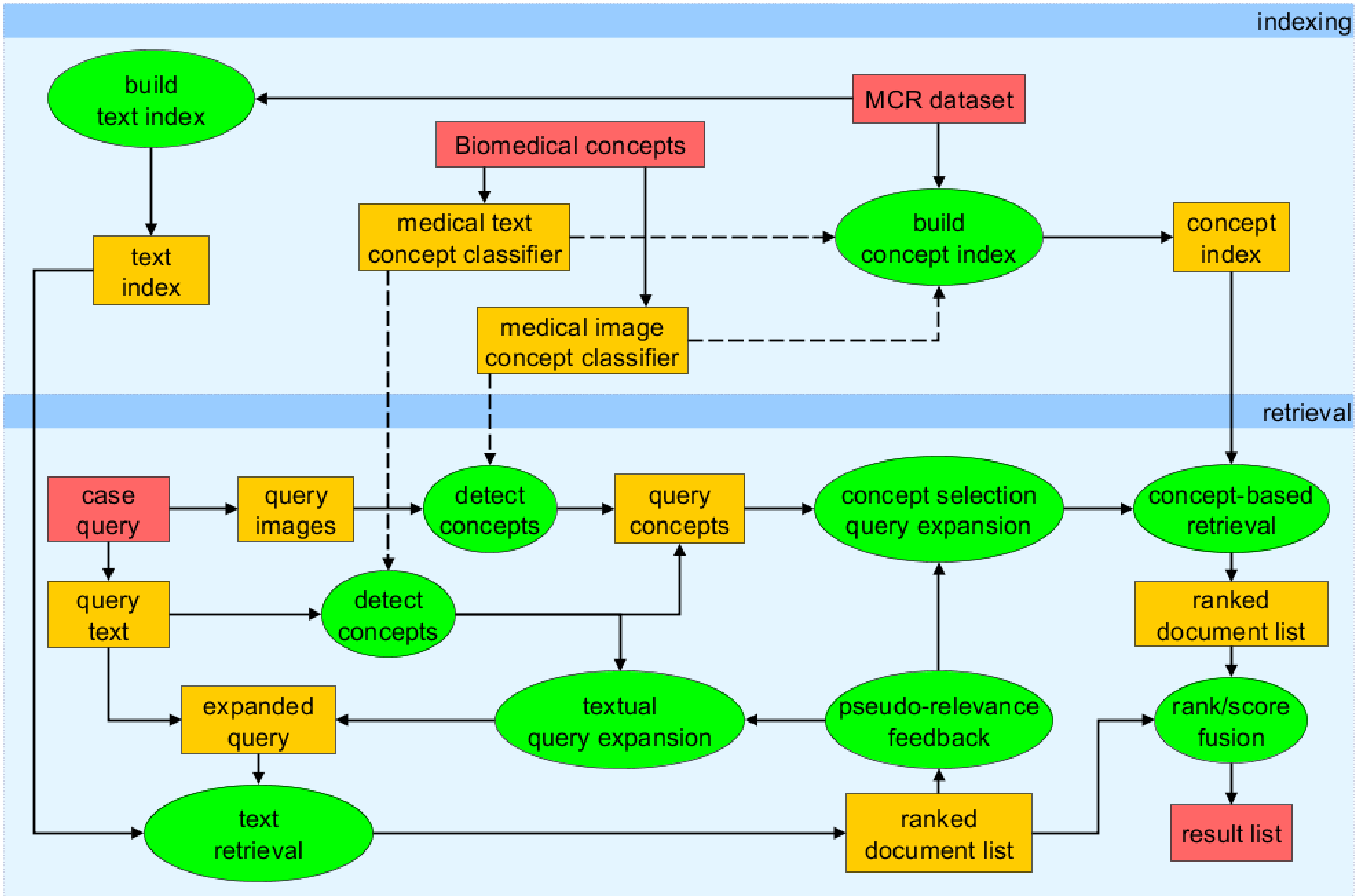


F	full text		+	MeSH document expansion
M	MeSH query expansion (BinDist)		r	pseudo-relevance feedback
B13	best result at ImageCLEF 2013			

OUTLINE 5

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



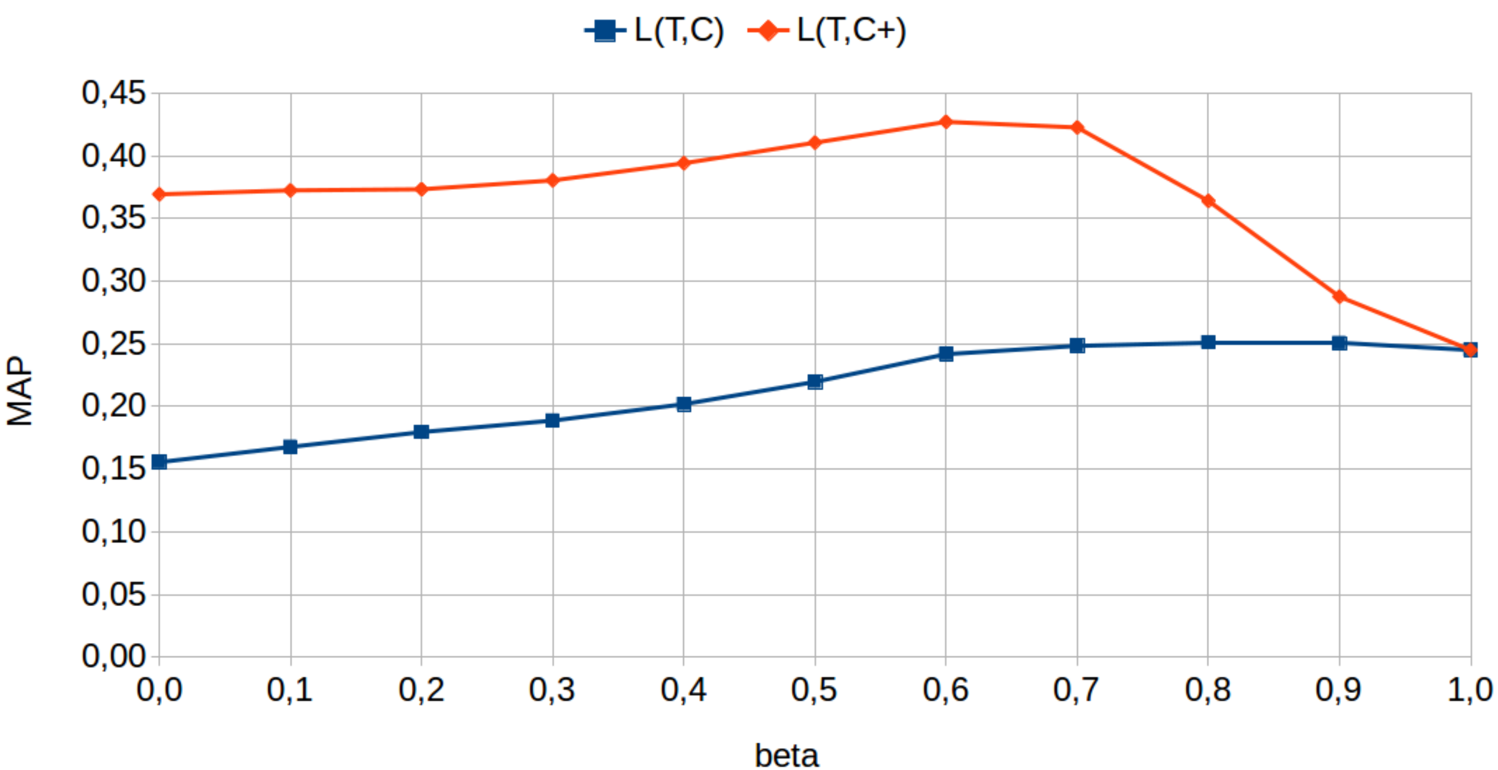
Fusion Methods

- Fuse result lists of retrieval methods A and B
- Linear fusion: $s = \beta * s_A + (1- \beta) * s_B$
  - Combine retrieval scores with fixed weight β
  - s_A, s_B: logistic score normalization from rank positions
- Query-adaptive fusion (QAF):
  - For each query q, choose weight β depending on q
  - E.g. by estimating performance of A and B for q

    $$\beta = p_A^2 / (p_A^2 + p_B^2)$$
  - "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 β)
C: concept-based retrieval (textual kNN concept mapping)
C+: concept-based retrieval with ground-truth MeSH terms

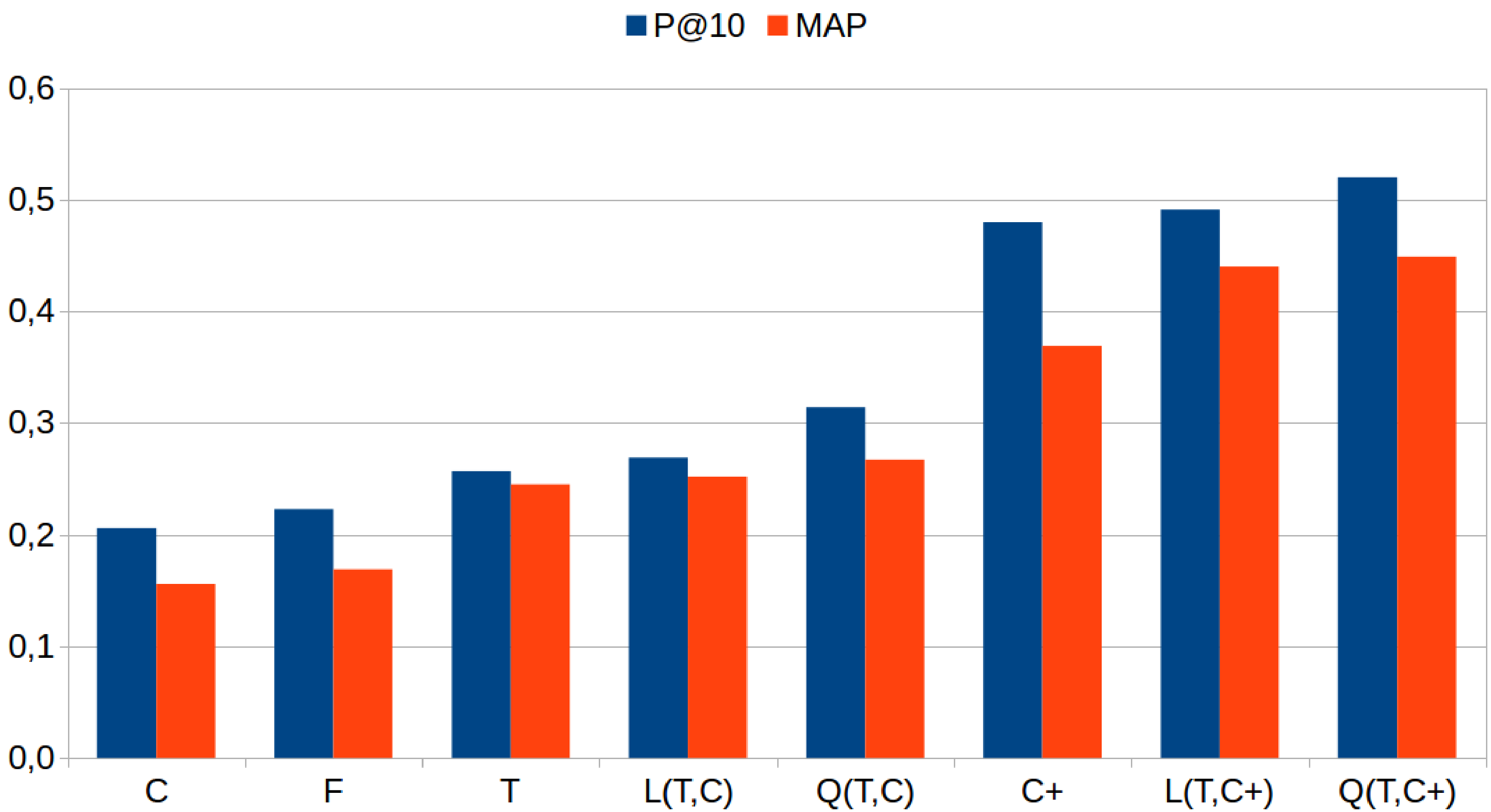


Fusion Results

L: linear fusion with optimized weight

Q: ideal query-adaptive fusion

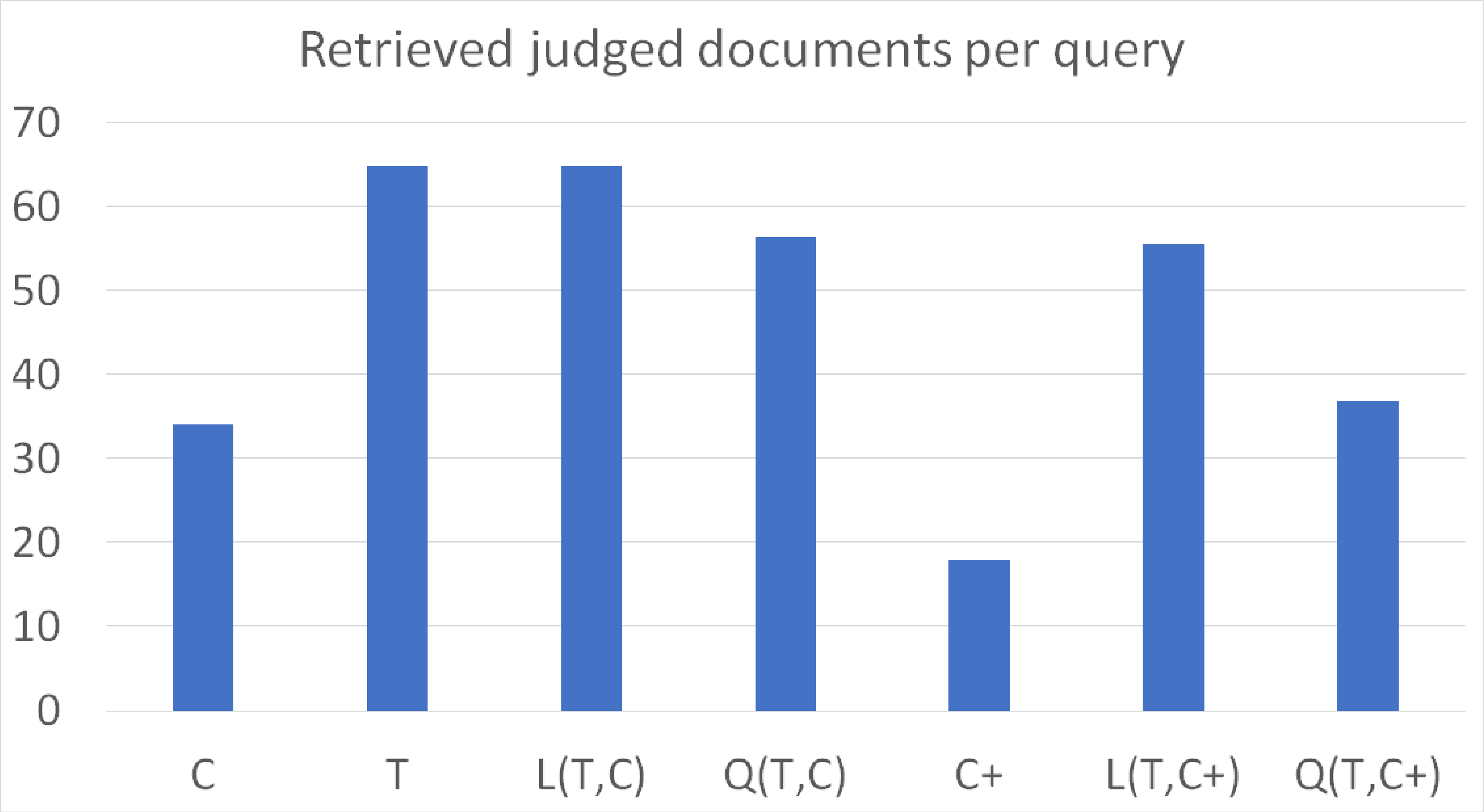
F: fulltext retrieval



OUTLINE 6

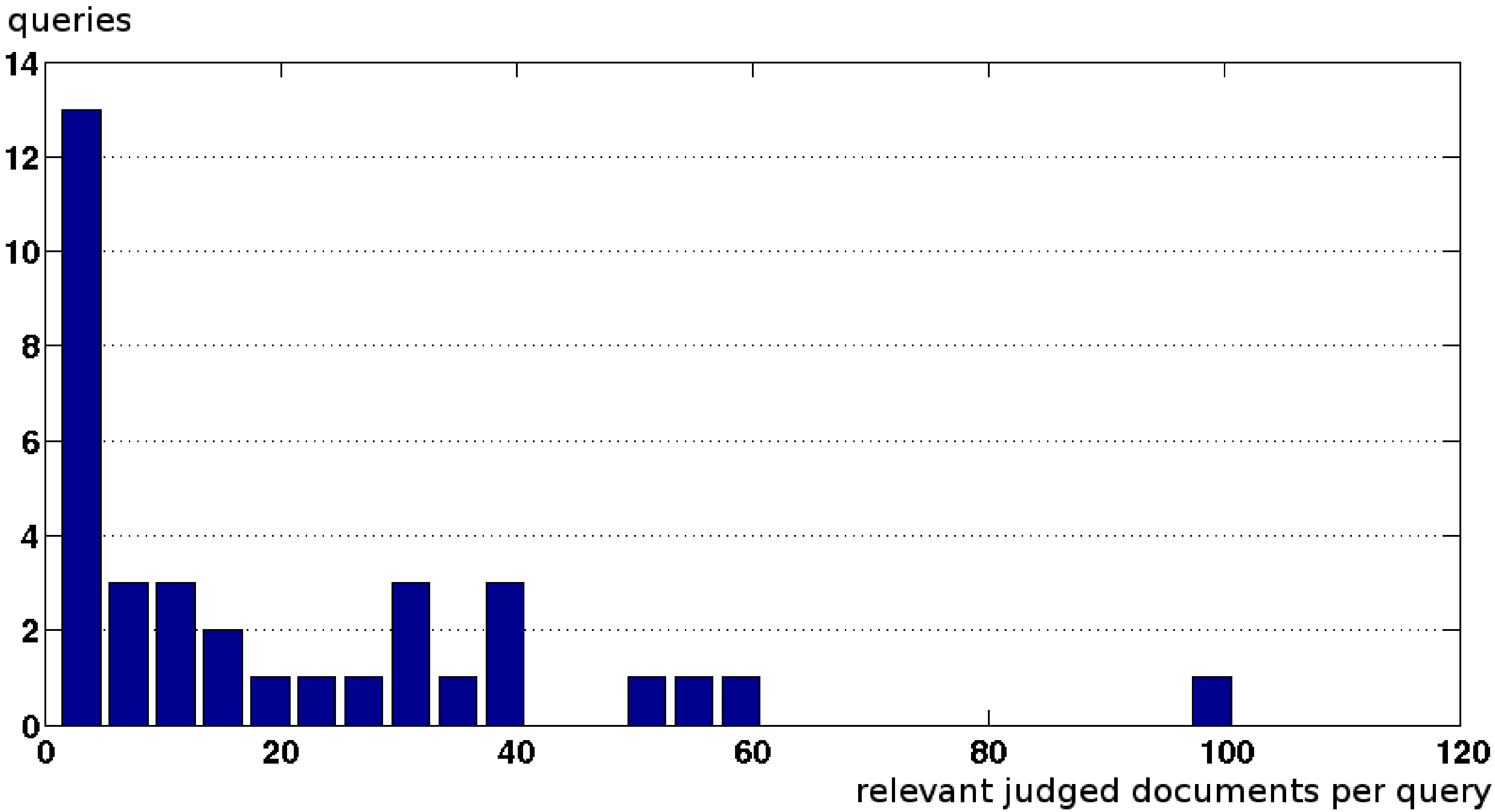
1. Introduction
2. Processing Compound Figures
3. Biomedical Concept Mapping
4. Using Concepts for textual MCR
5. Multimodal MCR
6. Limitations and Further Work

Limitations of MCR Dataset 1



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

[1] Mario Taschwer and Oge Marques. "Automatic separation of compound figures in scientific articles". In: *Multimedia Tools and Applications* (2016), pp. 1–30. ISSN: 1573-7721. DOI: 10.1007/s11042-016-4237-x.

[2] Mario Taschwer and Oge Marques. "Compound Figure Separation Combining Edge and Band Separator Detection". In: *MultiMedia Modeling*. Ed. by Qi Tian et al. Vol. 9516. Lecture Notes in Computer Science. Springer International Publishing, 2016, pp. 162–173. ISBN: 978-3-319-27670-0. DOI: 10.1007/978-3-319-27671-7_14.

[3] Mario Taschwer and Oge Marques. "AAUITEC at ImageCLEF 2015: Compound Figure Separation". In: *CLEF 2015 Working Notes*. Vol. 1391. CEUR Workshop Proceedings, ISSN 1613-0073. Toulouse, France, Sept. 2015. URL: http://ceur-ws.org/Vol-1391/25-CR.pdf.

PUBLICATIONS 2

[4] Mario Taschwer. "Medical Case Retrieval". In: *Proceedings of the 22nd ACM International Conference on Multimedia*. MM '14. Orlando, Florida, USA: ACM, 2014, pp. 639–642. ISBN: 978-1-4503-3063-3. DOI: 10.1145/2647868.2654856.

[5] Mario Taschwer. *Textual methods for medical case retrieval*. Tech. rep. TR/ITEC/14/2.01. Institute of Information Technology (ITEC), AAU Klagenfurt, Austria, May 2014. URL: http://www.itec.aau.at/bib/files/textual-mcr.pdf.

[6] Mario Taschwer. "Text-Based Medical Case Retrieval Using MeSH Ontology". In: *CLEF 2013 Evaluation Labs and Workshop, Online Working Notes*. Ed. by Pamela Forner, Roberto Navigli, and Dan Tufis. Valencia, Spain: CLEF Initiative, Sept. 2013, p. 5. ISBN: 978-88-904810-5-5. URL: http://ceur-ws.org/Vol-1179/CLEF2013wn-ImageCLEF-Taschwer2013.pdf.