Textual Methods for Medical Case Retrieval

Mario Taschwer

Institute of Information Technology (ITEC), Alpen-Adria Universität Klagenfurt, Austria

Problem: Medical Case Retrieval (MCR)

- Given a description of patient symptoms (query), find descriptions of diseases or patients' health records (document corpus) that are relevant as decided by medical experts.
- **How can text retrieval be improved for MCR?**

Novel MeSH Term Matching Algorithms

- MeSH (Medical Subject Headings) is a controlled vocabulary used to annotate biomedical publications.
- Novel algorithms to associate queries or documents with MeSH terms:

  t0 – BinCov binary coverage

  t1 – Dist distance-based match frequency

  t2 – BinDist combination of *BinCov* and *Dist* for matching runs

  t3 – IdfBinDist *BinDist* with score boosting by maximal IDF of MeSH term words

  t4 – IdfCovDist combination of *Dist* with IDF-based run coverage
- These methods are efficient and do not rely on natural language processing or machine learning.

Query and Document Expansion Methods

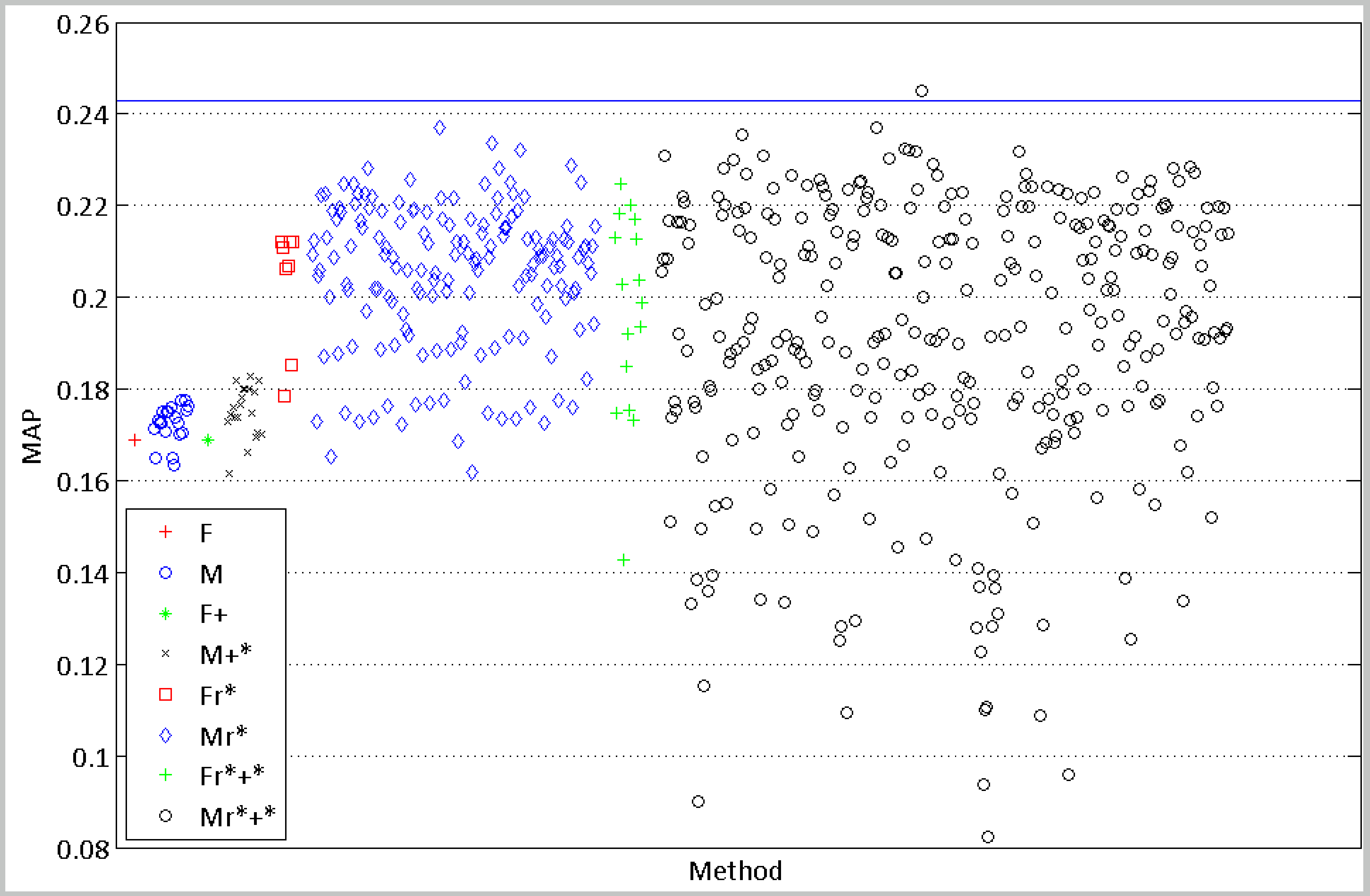
Acronym	Method	Count
F	*fulltext search* (no MeSH query expansion)	1
M	*MeSH query expansion*	20
tN	MeSH term matching algorithm, $\mathbf{0} \leq \mathbf{N} \leq \mathbf{4}$	5
xN	synonym selection method, $\mathbf{0} \leq \mathbf{N} \leq \mathbf{3}$	4
r*	*pseudo-relevance feedback*	8
r	unigrams ranked by TF-IDF	1
r2	unigrams and bigrams ranked by TF-IDF	1
rm	manually annotated MeSH terms	1
rm2	union of r and rm features	1
raN	automatically annotated MeSH terms ranked by score tN, $\mathbf{1} \leq \mathbf{N} \leq \mathbf{4}$	4
+*	*document expansion*	5
+	manually annotated MeSH terms	1
+N	automatically annotated MeSH terms ranked by score tN, $\mathbf{1} \leq \mathbf{N} \leq \mathbf{4}$	4

Parameter Optimization

Parameter	Type	Range	Description
$\mathbf{s_{min}}$	real	0.2 – 2.0	minimal matching score for MeSH term selection
μ_M	real	0.1 – 1.0	weighting factor of MeSH expansion terms relative to original query terms
m	integer	1 – 20	number of pseudo-relevant documents
k	integer	1 – 150	number of expansion terms to use for pseudo-relevance feedback
$\mathbf{k_2}$	integer	1 – 50	number of bigrams to use for expansion for **rf2** method
μ_F	real	0.1 – 2.0	weighting factor of feedback terms relative to original query terms
κ	real	0.1 – 2.0	relative importance of the two scoring functions for **rf2** and **rfm2** methods

- Each of the 546 evaluated method combinations (see scatterplot) was optimized for parameters on the ImageCLEF 2012 MCR dataset before evaluation on the 2013 dataset.

Evaluation on ImageCLEF 2013 MCR Dataset



Acronym	Group of methods	Count
F	fulltext search (without query expansion)	1
M	MeSH query expansion	20
F+	fulltext search with document expansion (manual MeSH annotation)	1
M+	MeSH query expansion with document expansion (manual MeSH annotation)	20
Fr*	fulltext search with pseudo-relevance feedback	8
Mr*	MeSH query expansion followed by pseudo-relevance feedback	160
Fr*+*	fulltext search with pseudo-relevance feedback and document expansion Fr+, Frm+, FraN+N, Frm2+*, Fr2+*	16
Mr*+*	MeSH query expansion followed by pseudo-relevance feedback with document expansion Mr+, Mrm+, MraN+N, Mrm2+*, Mr2+*	320
Total count		546

Conclusion

- Combination of MeSH query expansion and pseudo-relevance feedback substantially improves MCR performance over fulltext-only retrieval, achieving state-of-the-art effectiveness.
- Adding document expansion with MeSH terms does not provide additional benefit.
- There is no consistent best method within the set of proposed MeSH term matching algorithms.