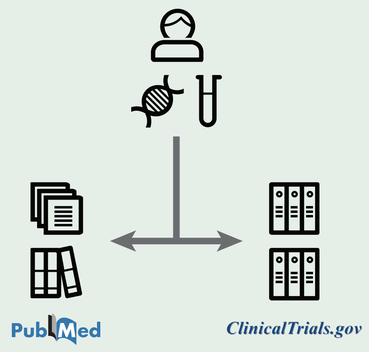
imi_mug at TREC 2017 Precision Medicine Track

Pablo López-García, Michel Oleynik, Zdenko Kasáč, Stefan Schulz
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz

Introduction

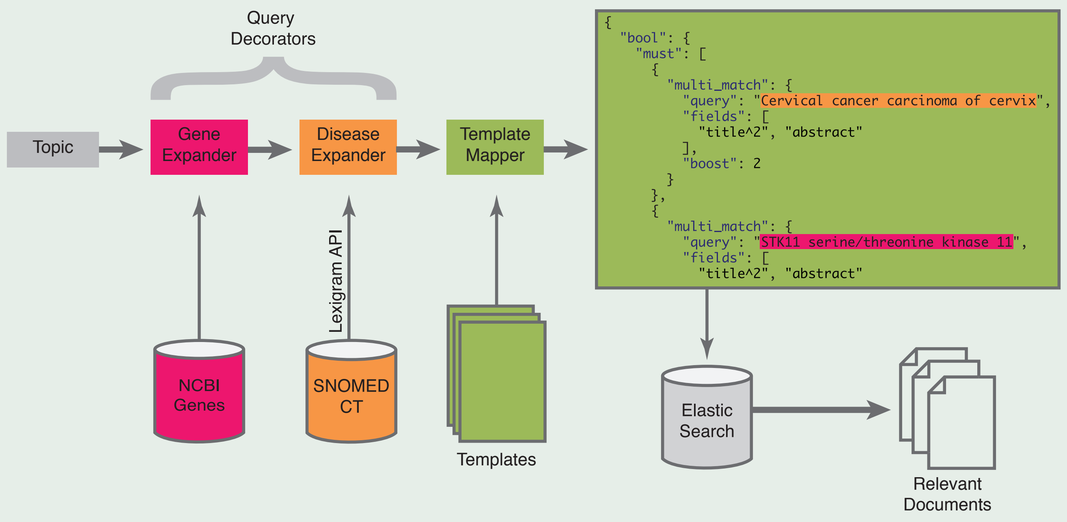
- 30 input topics: cancer patients
- Challenge: find relevant documents from two collections
  - Biomedical articles: Pubmed abstracts @ Jan 17 + ASCO/AACR proceedings
  - Clinical trials: ClinicalTrials.gov @ Apr 17

```
<topic number="15">
  <disease>Cervical cancer</disease>
  <gene>STK11</gene>
  <demographic>26-year-old female</demographic>
  <other>None</other>
</topic>
```



Infrastructure and Framework

- Elasticsearch 5.4.0 as search engine backend
- Open-source framework available on GitHub under the MIT License
  - Based on modular blocks: query templates and query decorators
- Reference standard built for the biomedical articles subtask
  - 739 topic-document pairs
  - Three annotators: two medical doctors, one computer scientist
  - 46 topics annotated by at least two annotators



Strategies and Resources

- Usage of `must` and `should` clauses in Elasticsearch
- Boosting of ASCO and AACR documents
  - Enabled for every submitted run
- Query expansion: diseases
  - Based on the Lexigram API [1]
  - Uses SNOMED CT, MeSH and ICD classification systems
- Query expansion: genes
  - Based on the NCBI Homo Sapiens Gene List [2]
  - Only expansion with the column `description` improved results
- Keyword boosting: positive and negative
  - N-grams extracted from the examples provided (Extra Topics)
  - Manual inspection of results during reference standard creation
  - Previous medical knowledge
- Positive keyword boosting: common chemotherapy suffixes
  - Extracted from the code L section of the ATC classification system

Positive boosters			Negative boosters	
surgery	resistance	-mab	transcript	probes
therapy	recurrence	-nib	paraffin	detection
treatment	targets	-cin	tumorigenesis	screening
prognosis	malignancy	-one	embedded	
prognostic	study	-ate	formalin	
survival	therapeutical	-mus	fish	
patient	outcome	-lin	tissue	

Table: Final list of positive and negative keyword boosters.

Some References

[1] Lexigram Inc. Lexigram HTTP API Documentation. https://docs.lexigram.io, 2017.

[2] National Center for Biotechnology Information (NCBI). NCBI Gene Lists. ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO, 2017.

Biomedical Articles: Submitted Runs

Strategy	mugpubboost	mugpubshould	mugpubbase	mugpubdiseas	mugpubgene
Matching disease + gene on MeSH and text fields	must	should	must	must	must
Keyword boosting: chemotherapy suffixes	Y	Y		Y	Y
Keyword boosting: positive	Y	Y	Y	Y	Y
Keyword boosting: negative	Y	Y		Y	Y
Query expansion: diseases				Y	
Query expansion: genes					Y
infNDCG	0.4088	0.2864	0.4031	0.3596	0.3016
P@10	0.6172	0.4483	0.6138	0.4759	0.4552
R-Prec	0.2735	0.1887	0.2743	0.2393	0.2065

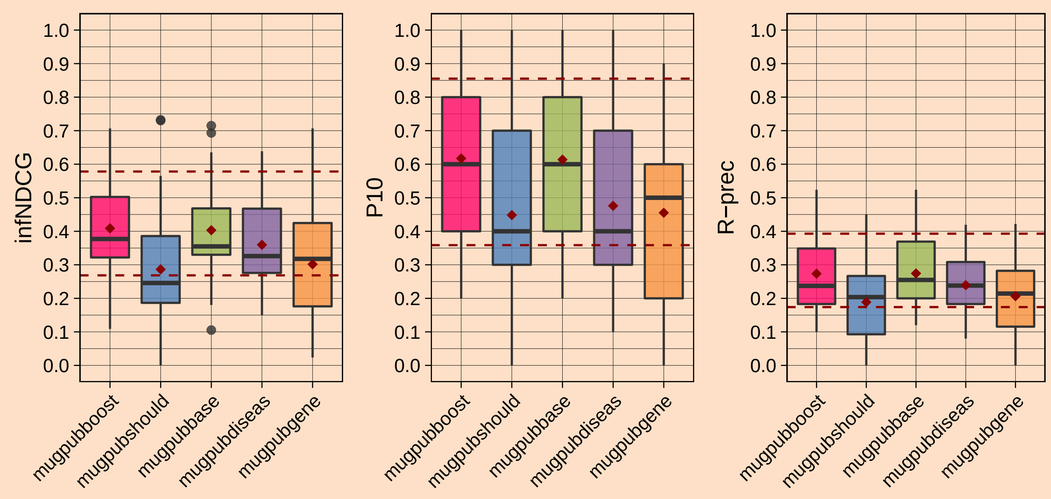


Figure: Boxplots comparing different runs to the overall best and median results.

Clinical Trials: Submitted Runs

Strategy	mugctboost	mugctdisease	mugctbase	mugctgene	mugctmust
Matching disease + gene on title and inclusion criteria	must	must	should	must	should
Matching age + sex on metadata	must	must	must	must	should
Comorbidities should not match exclusion criteria	Y	Y		Y	Y
Keyword boosting: positive	Y	Y		Y	
Query expansion: diseases		Y			
Query expansion: genes				Y	
P@5	0.2500	0.0929	0.3000	0.2357	0.2929
P@10	0.2286	0.0679	0.2643	0.2321	0.2536
P@15	0.1952	0.0738	0.2262	0.2119	0.2143

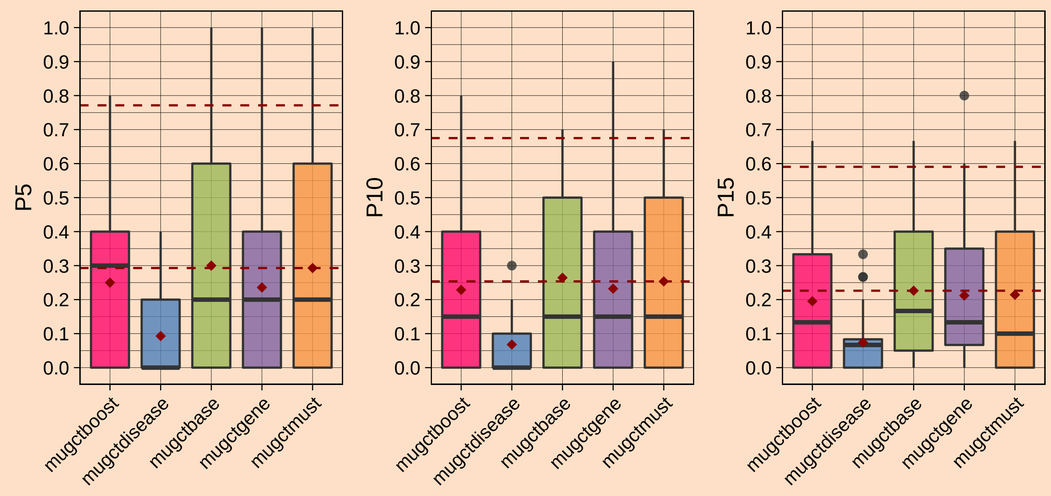


Figure: Boxplots comparing different runs to the overall best and median results.

Limitations and Future Work

- Improve recall
  - Synonym, hypo-, and hypernym list based on MeSH
  - Tiered indexes
  - Pseudo-relevance feedback
- Explore other ranking strategies
  - Use decay functions to push down older documents
  - Restrict to MeSH major topics
  - Debug Elasticsearch's `english` analyzer
  - Disable field-length normalization (aka BM15)

Acknowledgements

We thank Lexigram, Inc. for providing us with an API key to access their medical knowledge graph. Our work is partially funded by the Brazilian National Research Council - CNPq (project number 206892/2014-4).
Source code available at `https://goo.gl/axTw5J`.