Initial Experiments with Learning to Rank

QUT ielab at CLEF eHealth 2017 Technology Assisted Reviews Track

Harrisen Scells, Guido Zuccon, Anthony Deacon, Bevan Koopman

TAR Task

- Two tasks:

1. Produce an efficient ordering of studies retrieved by a boolean search strategy such that all of the relevant abstracts are retrieved as early as possible

2. Identify a subset of the ranked studies which contains all or as many of the relevant abstracts for the least effort

Our Approach

- We train **learning to rank** models using **domain specific features**
    - *What effect do PICO features have with respect to learning to rank algorithms?*
- Compare the PICO models to non-PICO models

PICO (Population, Intervention, Control or comparison and Outcome) is a technique used in evidence based practice to frame and answer clinical questions and is used extensively in the compilation of systematic reviews

What does this mean?

BM25 to rank

- Boolean “baseline” system simulating PubMed, others
- Replicate system that uses PICO

Learning to ranking

- Several L2R models trained on non-PICO
- Several L2R models not train on PICO

Interested in the effectiveness of the features and their effectiveness in L2R

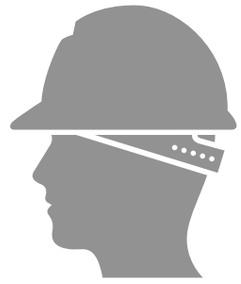
PICO Annotating

- PICO features were annotated on both the **document** and **query**
    - Document features extracted **automatically** using RobotReviewer [1]
    - Query features extracted **manually** with the assistance from a medical professional
        - The queries are the actual “search strategies” - **not the title of the topic**

[1] B C Wallace, J Kuiper, A Sharma, M B Zhu, and I J Marshall. 2016. Extracting PICO sentences from clinical trial reports using supervised distant supervision. JMLR 17, 132 (2016), 1–25.

What were the features?

- IDFSum - sum of the IDF scores
- IDFStd - std.dev. of the IDF scores



(feature engineer)

- IDFMax - max IDF score
- IDFAvg - mean IDF score
- PopulationCount, InterventionCount, OutcomeCount
  - Number of P,I,O terms overlapping in query and doc

What were the models?

- Various L2R models trained:
  - MART
  - LambdaMART
  - AdaRank
  - **Coordinate Ascent**
  - **Random Forests**

Submitted Runs

- Each model evaluated on an existing collection [2]
  - Also contains annotated documents and queries using the same method previously described

[2] Scells, Harrisen, et al. "A Test Collection for Evaluating Retrieval of Studies for Inclusion in Systematic Reviews." Proceedings of the 40th International ACM SIGIR (2017).

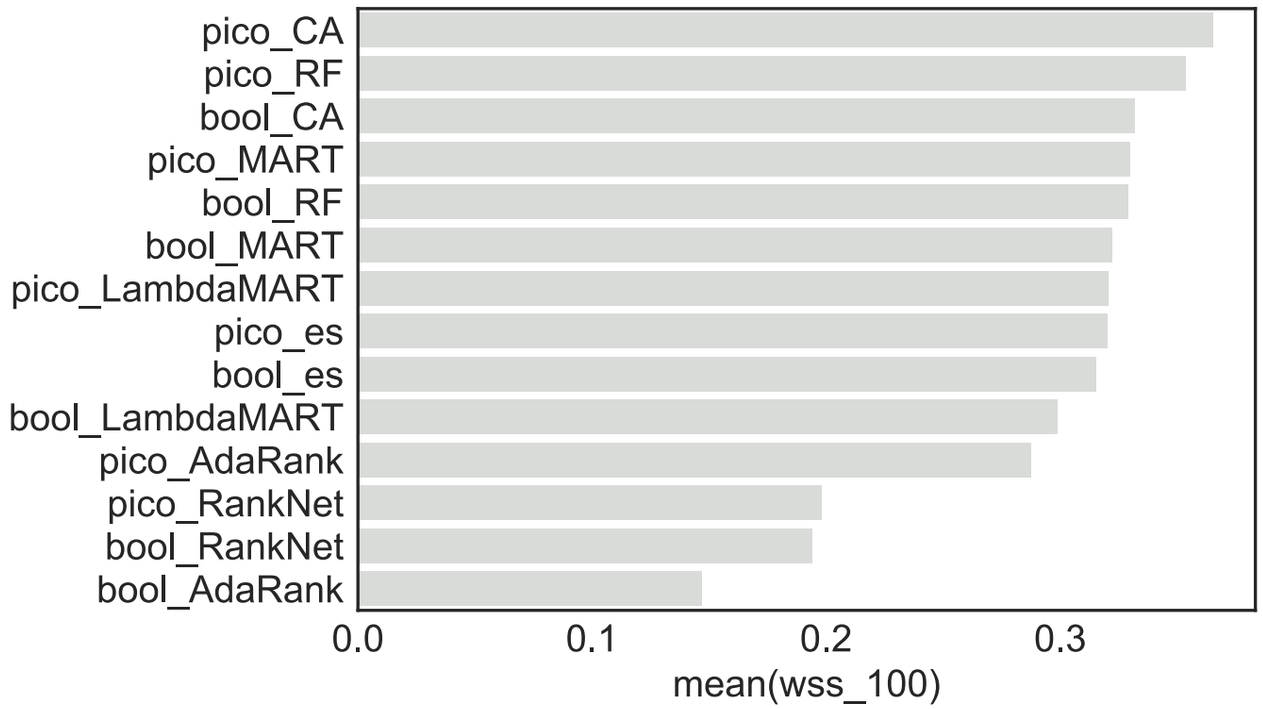
Model Evaluation

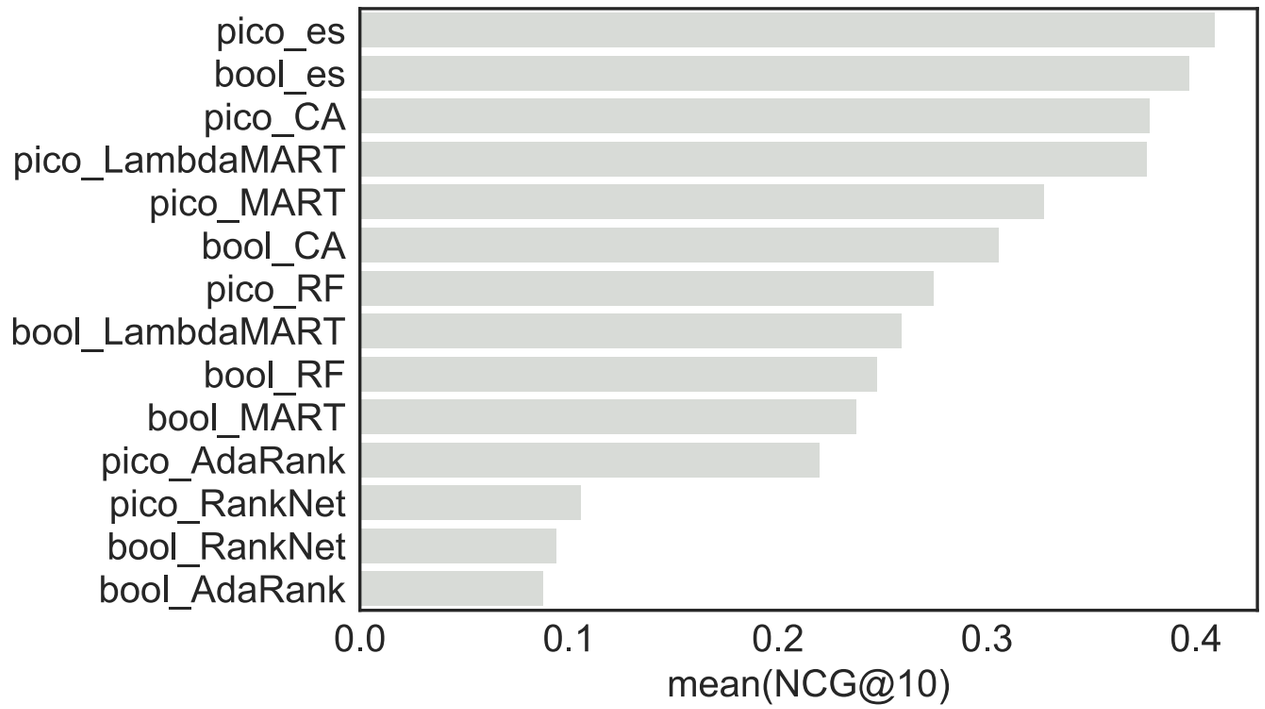
	NCG@10		AP	
	Boolean	PICO	Boolean	PICO
Elasticsearch	**0.397**	**0.409**	**0.104**	0.102
MART	0.237	0.327	0.066	0.086
AdaRank	0.0875	0.2197	0.0255	0.0619
Coordinate Ascent*	0.305	0.378	0.076	**0.114**
LambdaMART	0.259	0.377	0.068	0.097
Random Forests*	0.247	0.275	0.061	0.088

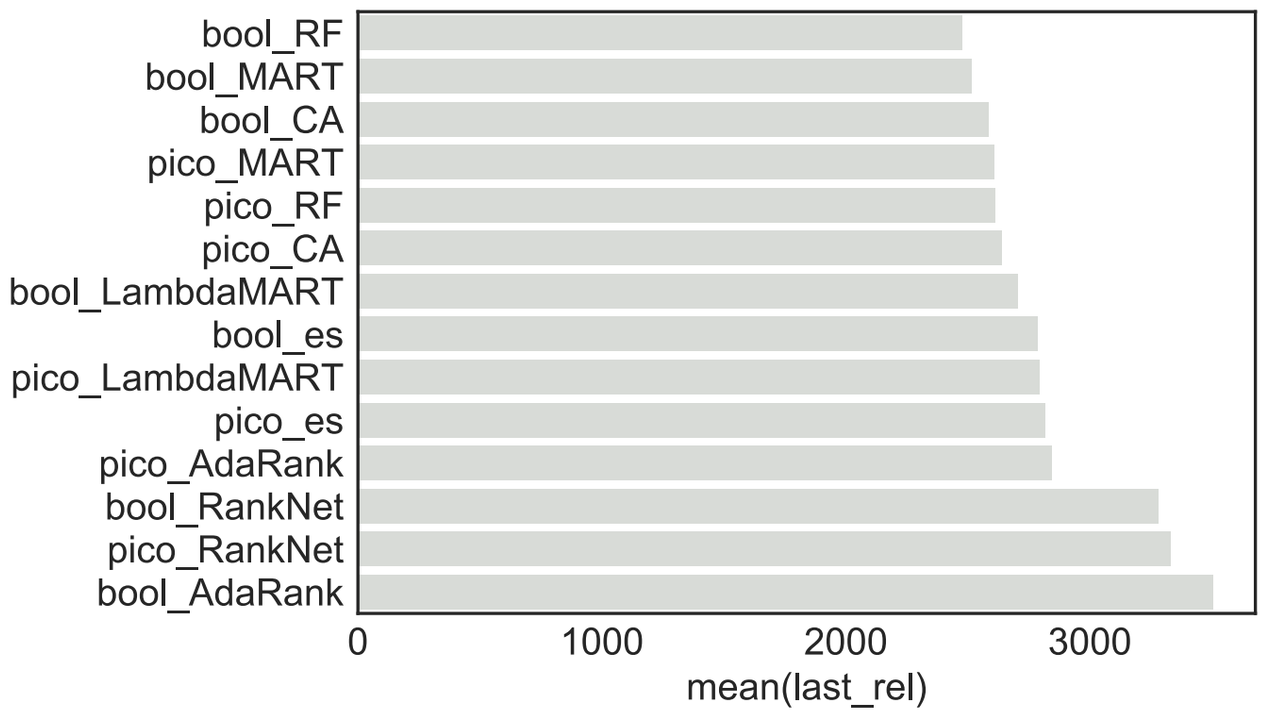
*submitted runs

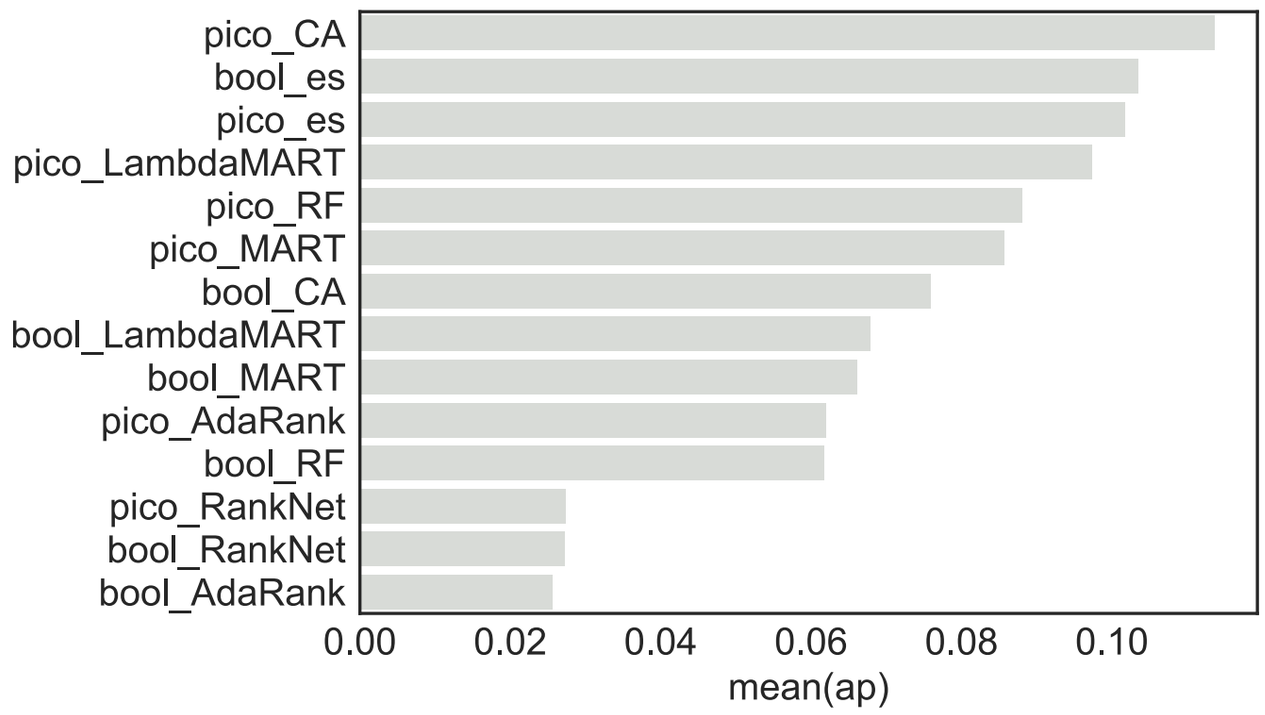
Did I get lucky/unlucky training these models?
How do I explain to medical researchers the reason they are seeing this ranking?
Small set of features - hard to beat Elasticsearch BM25

Results Breakdown









Questions


harrisen.scells@hdr.qut.edu


@hscells


github.com/hscells