Cumulative Citation Recommendation: A Feature-aware Comparisons of Approaches

Gebrekirstos G. Gebremeskel, Jiyin He*, Arjen P. de Vries*, and Jimmy Lin@

**CWI, Amsterdam, The Netherlands*

@University of Maryland, College Park

DEXA-TIR 2014, Munich, Germany

September 04, 2014

- IR
  - Machine learning
  - Uses many features
- Several studies
  - Compare IR approaches
  - Recommend best approach
- Scenario
  - Two methods (mtd 1 and mtd 2) and two feature sets (ftr 1 and ftr 2). Using ftr 1, mtd 1 outperforms mtd 2.

Question

- Does that mean mtd 1 will still outperform mtd 2 if we use ftr 2 in place of ftr 1?

Problem setting and dataset

- Problem setting:
  - Cumulative citation recommendation (CCR)
    - filtering a stream to identify those documents that are citation-worthy to Knowledge Base(KB) entities of interest
- Dataset:
  - TREC KBA-CCR-2012 dataset
    - 29 Wikipedia entities
    - Time-stamped stream of documents of news, social media content.
    - Relevance judgments for training and testing

State of the art CCR

- Three different approaches
  - String matching
  - Classification
    - Random Forest (CL-RF)
  - Learning to Rank (LTR)
    - Random forest (LTR-RF

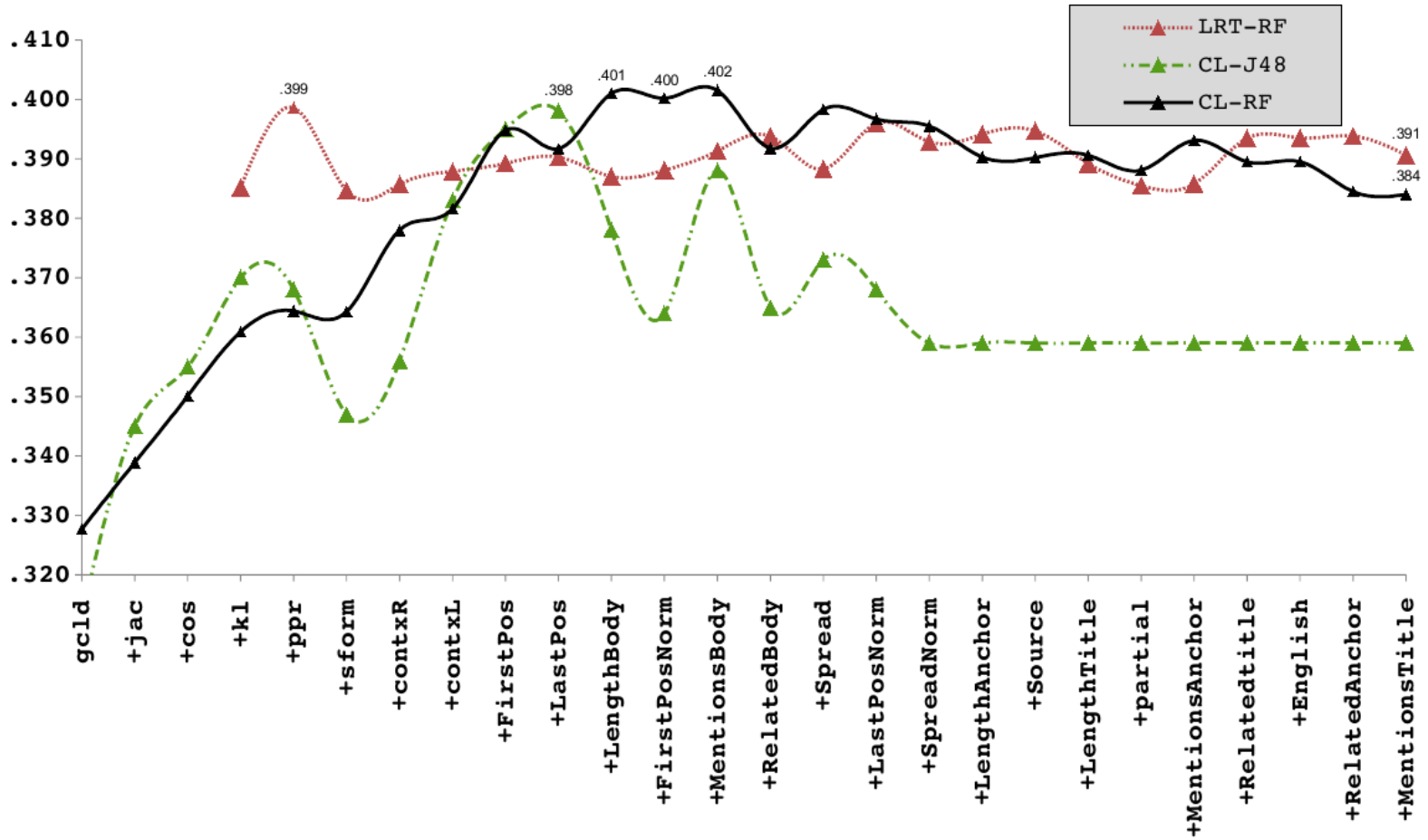
- LTR approaches outperform classification approaches for CCR task

A two-step approach

- Filtering using DBpedia name variants
- Subsequent classification or learning to rank

Feature selection

- Preliminary feature elimination
- Forward selection


LRT-RF
CL-J48
CL-RF
.410
.400
.390
.380
.370
.360
.350
.340
.330
.320
.399
.398
.401
.400
.402
.391
.384
gcld
+jac
+cos
+kl
+ppr
+sform
+contxR
+contxL
+FirstPos
+LastPos
+LengthBody
+FirstPosNorm
+MentionsBody
+RelatedBody
+Spread
+LastPosNorm
+SpreadNorm
+LengthAnchor
+Source
+LengthTitle
+partial
+MentionsAnchor
+Relatedtitle
+English
+RelatedAnchor
+MentionsTitle

Best Scores

Method	F	SU
MC-RF	0.360	0.263
MC-LTR-RF	**0.390**	**0.369**
LRE-KBA	0.377	0.329
CL-RF	**0.402**	**0.396**
LTRE-RF	0.394	0.411
CL-J48	0.388	0.306

Summary

- Identified a few, but more powerful features
- Fair comparison of several approaches from previous studies
- Found out that classification approaches outperform learning-to-rank approaches
  - contrary to previous findings

Take home ..

- Comparing approaches is problematic due to the interplay between the approaches themselves and the feature sets one chooses to use.