Ranking Health Web Pages with Relevance and Understandability

João Palotti[1], Lorraine Goeuriot[2], Guido Zuccon[3], Allan Hanbury[1]

[1]Vienna University of Technology, Austria - {palotti,hanburry}@ifs.tuwien.ac.at
[2]Universite Grenoble Alpes, France - lorraine.goeuriot@imag.fr
[3]Queensland University of Technology, Australia - g.zuccon@qut.edu.au

What is the Problem?

Acute myocardial infarction occurs due to coronary artery disease caused by a rupture of an atherosclerotic plaque.

VS

A heart attach happens when a blood vessel in the heart gets blocked and blood cannot get to part of the heart.

Aim

Devise method to combine both topical relevance and understandability for retrieval

Approach

Learning to rank exploiting topical and readability (proxy as understandability) features

DataSet

CLEF eHealth 2015 IR task dataset:
- 1 million health related web pages
- 67 queries
- 8,700 judgdments of topical and understandability relevance
- *https://github.com/CLEFeHealth*

Learning to rank Approach

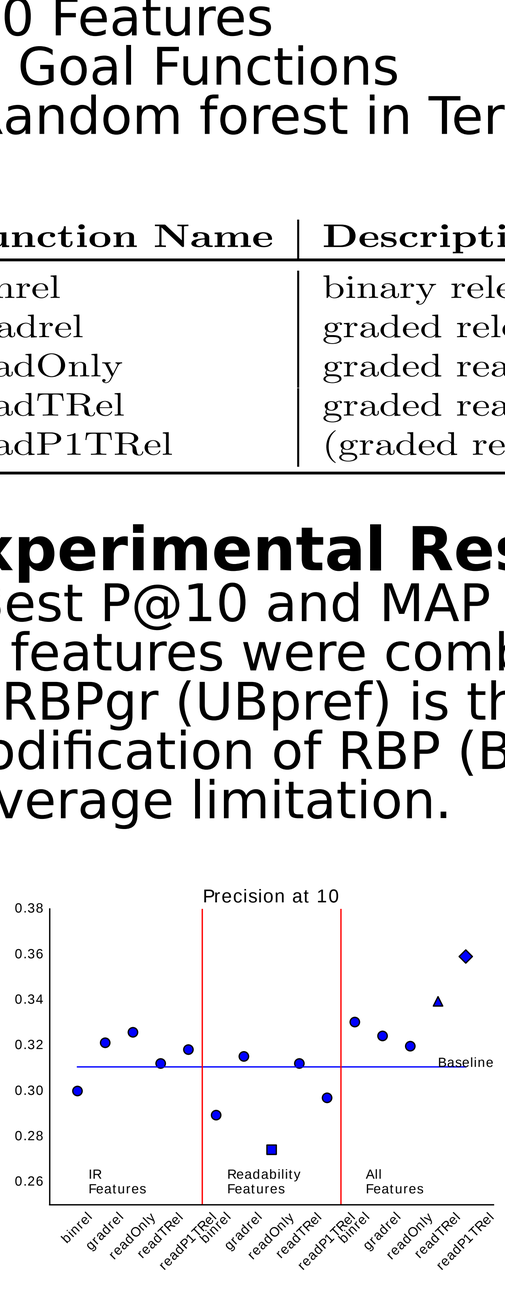
- 90 Features
- 5 Goal Functions
- Random forest in Terrier 4.0

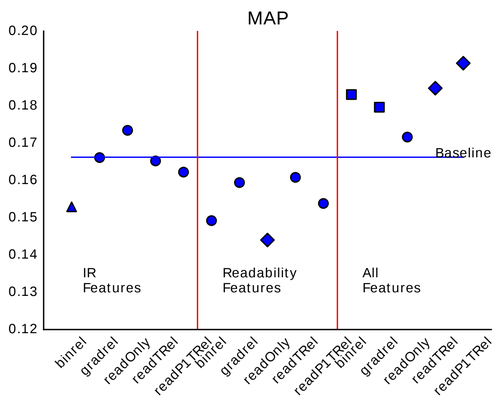
Function Name	Description
binrel	binary relevance only
gradrel	graded relevance only
readOnly	graded readability only
readTRel	graded readability × graded relevance
readP1TRel	(graded readability + 1) × graded relevance

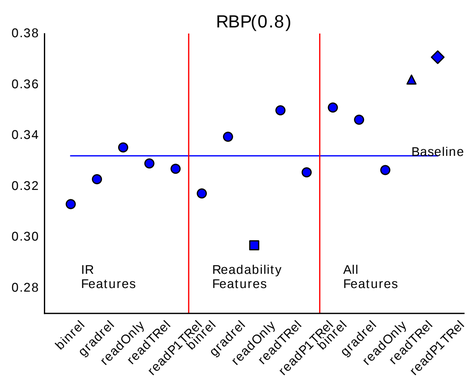
Feature Type	Feature Category	Feature Name
IR Features (18)	Common IR Models (14)	BM25 ⋆ PL2 ⋆ DirichletLM ⋆ LemurTF_IDF ⋆ TF_IDF ⋆ DFRee ⋆ Hiemstra_LM ⋆
	Query Independ. (2)	Document Length ⋆
	Doc. Score Modifier (2)	Divergence from Randomness Markov Random Field
Readability Features (72)	Traditional Formulas (8)	ARI Index Coleman Liau Index Dale-Chall Score Flesch Kincaid Grade Flesch Reading Ease Gunning Fog Index LIX Index SMOG Index
	Surface Measures (25)	# Characters ◇† # Sentences ◇ # Syllables ◇† # Words † # (|Syllables(Word)| >3) ◇† # (|Word| >4) ◇† # (|Word| >6) ◇† # (|Word| >10) ◇† # (|Word| >13) ◇†
	General Vocabulary Related Features (12)	Numbers ◇† English Dictionary ◇† Dale-Chall List ◇† stopwords ◇†
	Medical Vocabulary Related Features (27)	Acronyms ◇† Mesh ◇† DrugBank ◇† ICD10 (Inter. class. of Diseases) ◇† Medical Prefixes ◇† Medical Suffixes ◇† Consumer Health Vocabulary ◇† Sum(chv Score) ◇† Mean(chv Score) ◇†

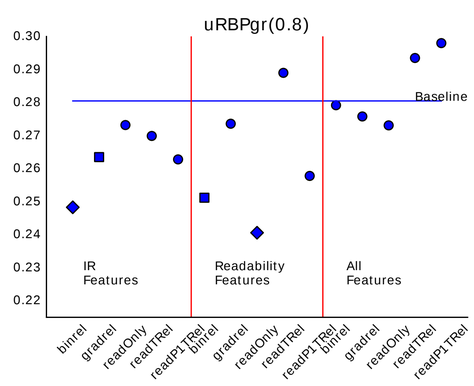
Experimental Results

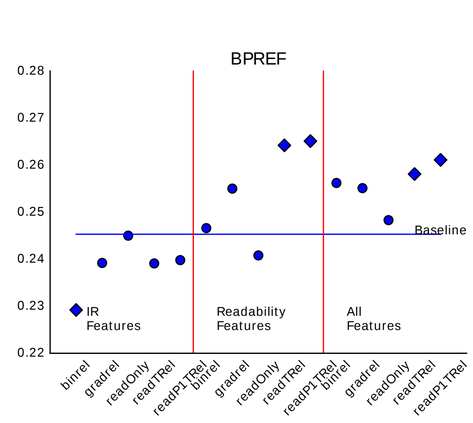
- Best P@10 and MAP were obtained when all features were combined.
- uRBPgr (UBpref) is the understandability modification of RBP (BPref): overcome coverage limitation.

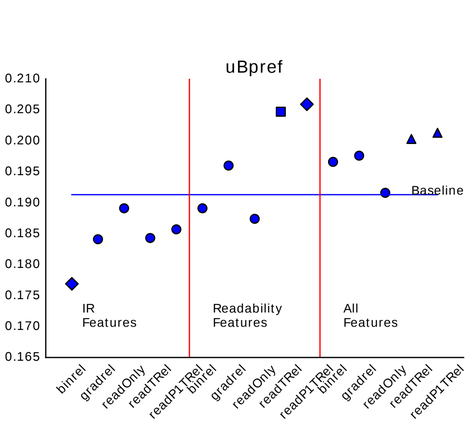












Conclusion

Incorporating readability features improves both traditional and understandability metrics. Stay tuned with CLEF eHealth 2016 (https://sites.google.com/site/clefehealth) and future!

See more results at: https://github.com/ielab/sigir2016-ranking-relevance-understandability