Evaluating Evaluation Measure Stability

by Chris Buckley and Ellen M. Voorhees

Presenters:

İzzeddin Gür

Mehmet Güvercin

Introduction

- The accuracy of evaluation measures in IR
- Three rules-of-thumb
  - Reasonable number of requests
    - TREC used 25 as minimum, 50 as norm
    - Jones and Rijsbergen suggested 75 as minimum
  - Reasonable evaluation measure
    - Avg Precision, R-Precision and Precision@20(or 10, 30)
  - Reasonable notion of difference
    - Jones and Rijsbergen suggested 5% as noticeable, 10% is material

Motivation and Purpose

- Very little attention has been paid to explore how the properties of these rules-of-thumb support conclusions as to whether one retrieval method is better than other.

- Experimentally quantifying the likely error associated with the conclusion “method A is better than method B” given a number of requests, an evaluation measure, and a notion of difference

Computing error rate

$$ErrorRate = \frac{\sum Min(|A > B|, |B > A|)}{\sum(|A > B| + |A < B| + |A == B|)}$$

	INQa			INQe			INQp			Saba			Sabe			Sabm			acs			pir		
APL	18	0	3	2	11	8	19	0	2	11	0	10	0	19	2	3	11	7	21	0	0	0	19	2
INQa				0	21	0	4	6	11	0	14	7	0	21	0	0	21	0	21	0	0	0	21	0
INQe							21	0	0	19	0	2	1	16	4	4	4	13	21	0	0	0	17	4
INQp										0	15	6	0	21	0	0	21	0	21	0	0	0	21	0
Saba													0	21	0	0	21	0	21	0	0	0	21	0
Sabe																21	0	0	21	0	0	2	4	15
Sabm																			21	0	0	0	19	2
acs																						0	21	0

a) Average Precision

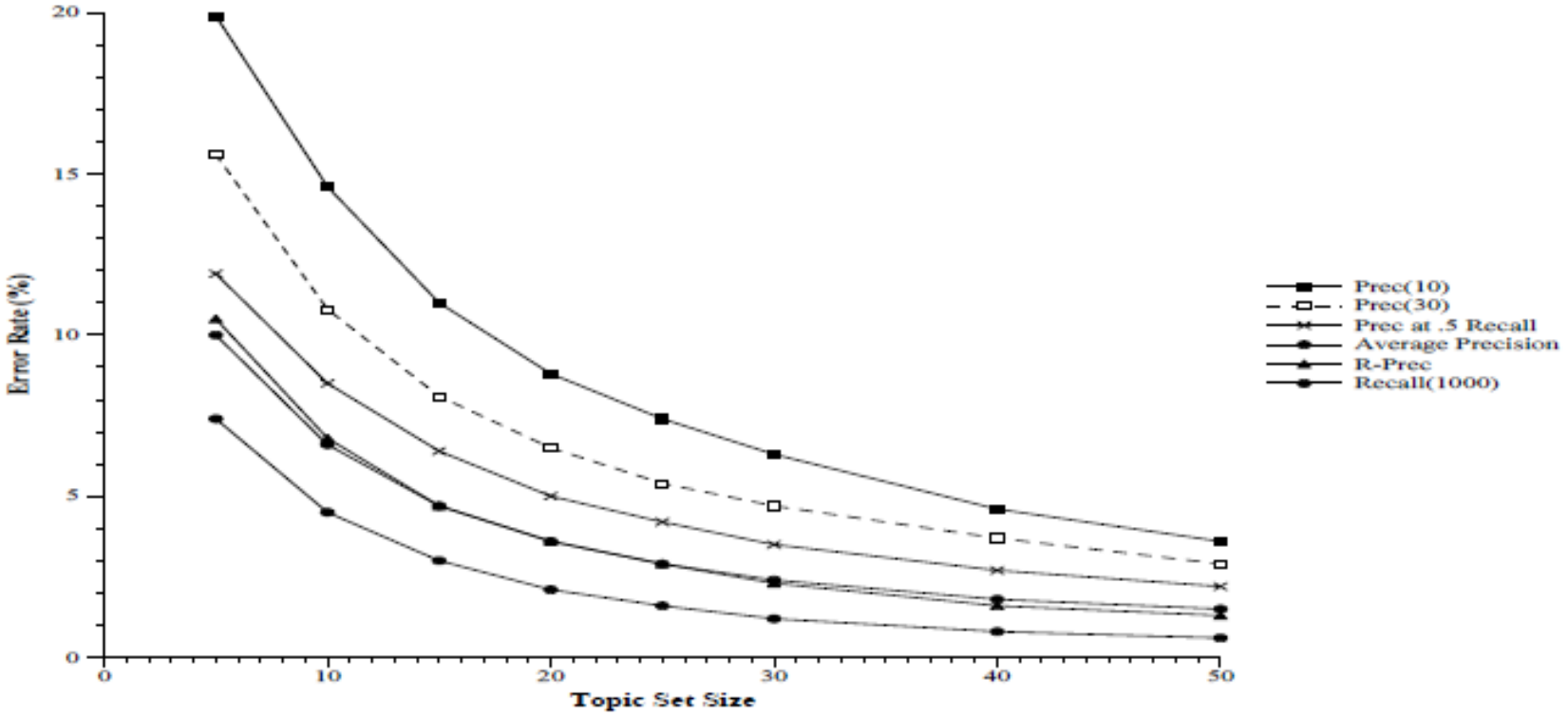
	INQa			INQe			INQp			Saba			Sabe			Sabm			acs			pir		
APL	2	12	7	0	19	2	3	9	9	2	11	8	0	20	1	1	14	6	13	1	7	0	19	2
INQa				0	14	7	4	2	15	2	6	13	0	21	0	0	9	12	18	0	3	0	15	6
INQe							20	0	1	16	1	4	4	6	11	14	2	5	21	0	0	6	4	11
INQp										2	5	14	0	20	1	1	12	8	18	0	3	0	19	2
Saba													0	19	2	0	6	15	17	0	4	0	16	5
Sabe																18	0	3	21	0	0	8	1	12
Sabm																			19	0	2	1	12	8
acs																						0	21	0

b) Prec(10)

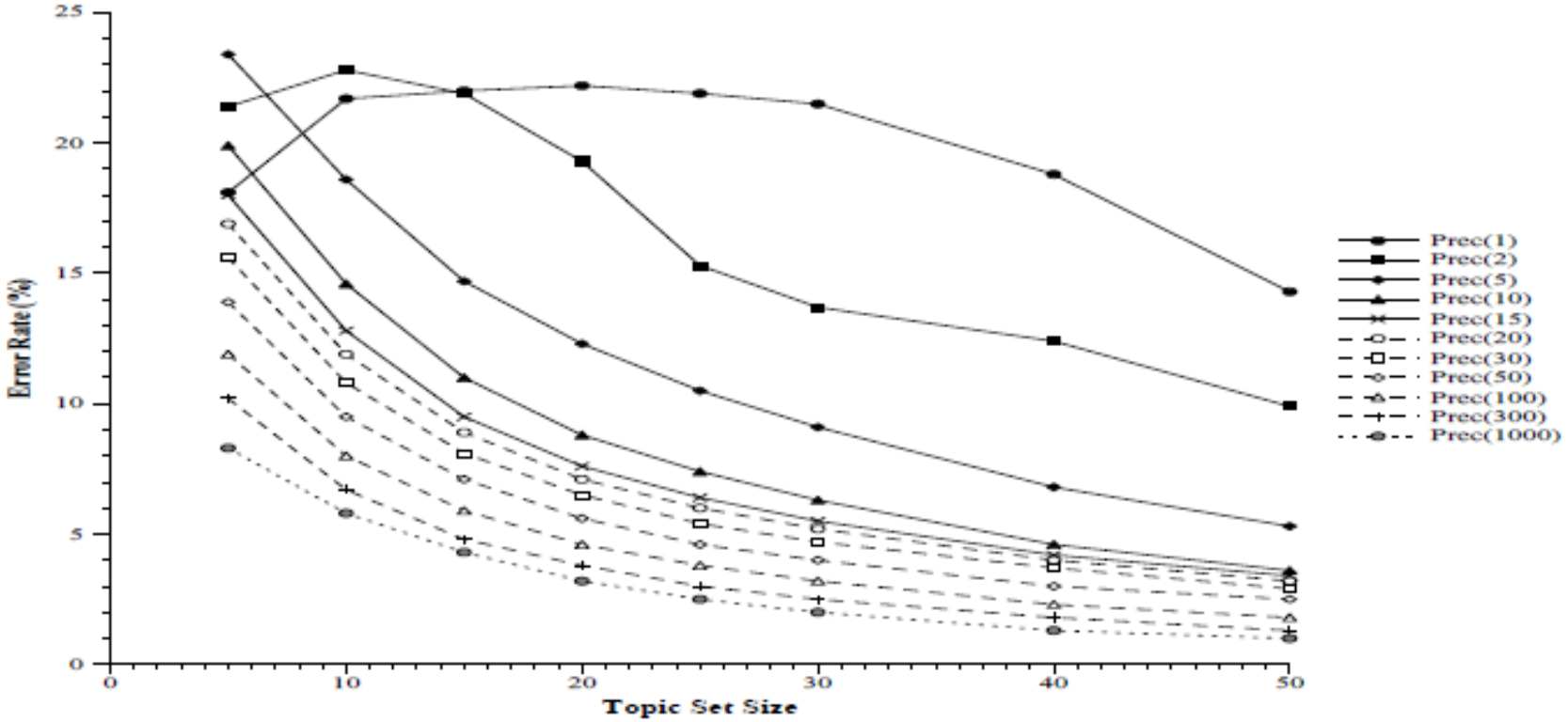
A>B, B>A, A=B values for different pairs of retrieval methods

Measure	Error Rate (%)	Std. Dev. (%)	Ties (%)
Prec(1)	14.3	1.3	23.4
Prec(10)	3.6	0.9	24.3
Prec(30)	2.9	0.8	23.8
Prec at .5 R	2.2	0.5	11.4
Prec(100)	1.8	0.5	20.7
Ave Prec	1.5	0.4	12.8
R-Prec	1.3	0.4	19.1
Prec(1000)	1.0	0.4	22.5
Recall(1000)	0.6	0.2	20.8

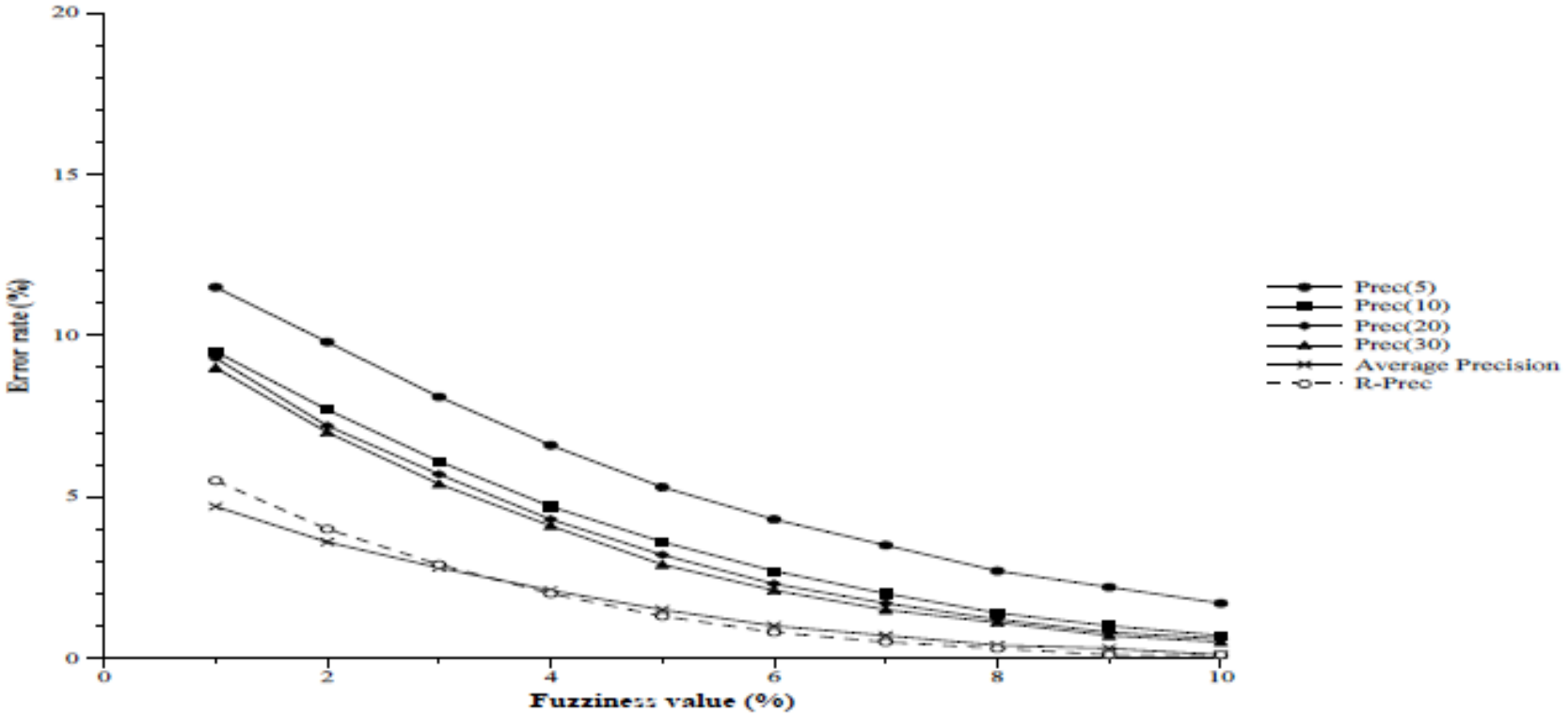
Error rates, std deviations and percentage of ties



Error rate according to varying topic set size



Error rates of Prec(λ) values according to varying topic set size



Error rates of different measures according to varying fuzziness value

Conclusion

- #Requests
  - Error rate increases as the number of requests decreases
- Evaluation of Measures
  - Prec (λ) is less stable, except Prec(1000)
  - R-Precision and Avg. Precision have similar results but Avg. Precision has more discriminative power
  - Recall(1000) is very stable but only appropriate if finding all relevant documents is important
  - For general purpose retrieval, Avg Precision is suitable
  - If #relevant documents is unknown, Prec(20), Prec(30) is suitable
- Notion of Difference
  - Larger difference threshold decreases error rate, however with cost of decreasing discrimination power.

Q & A