Evaluating Evaluation Measure Stability

by Chris Buckley and Ellen M. Voorhees

presenters: Sefa Kılıç and Anıl Türel

April 15, 2010

Introduction

- Three rules-of-thumb
  - Reasonable number of queries.
  - Reasonable evaluation measure.
    - Average Precision
    - R-Precision
    - Precision(20).
  - Threshold for method comparison.
- The paper examines these three rules-of-thumb.

Error Rates for a Variety of Measures

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

Figure: Error rate was computed using a fuzziness factor of 5%.

Varying topic set size

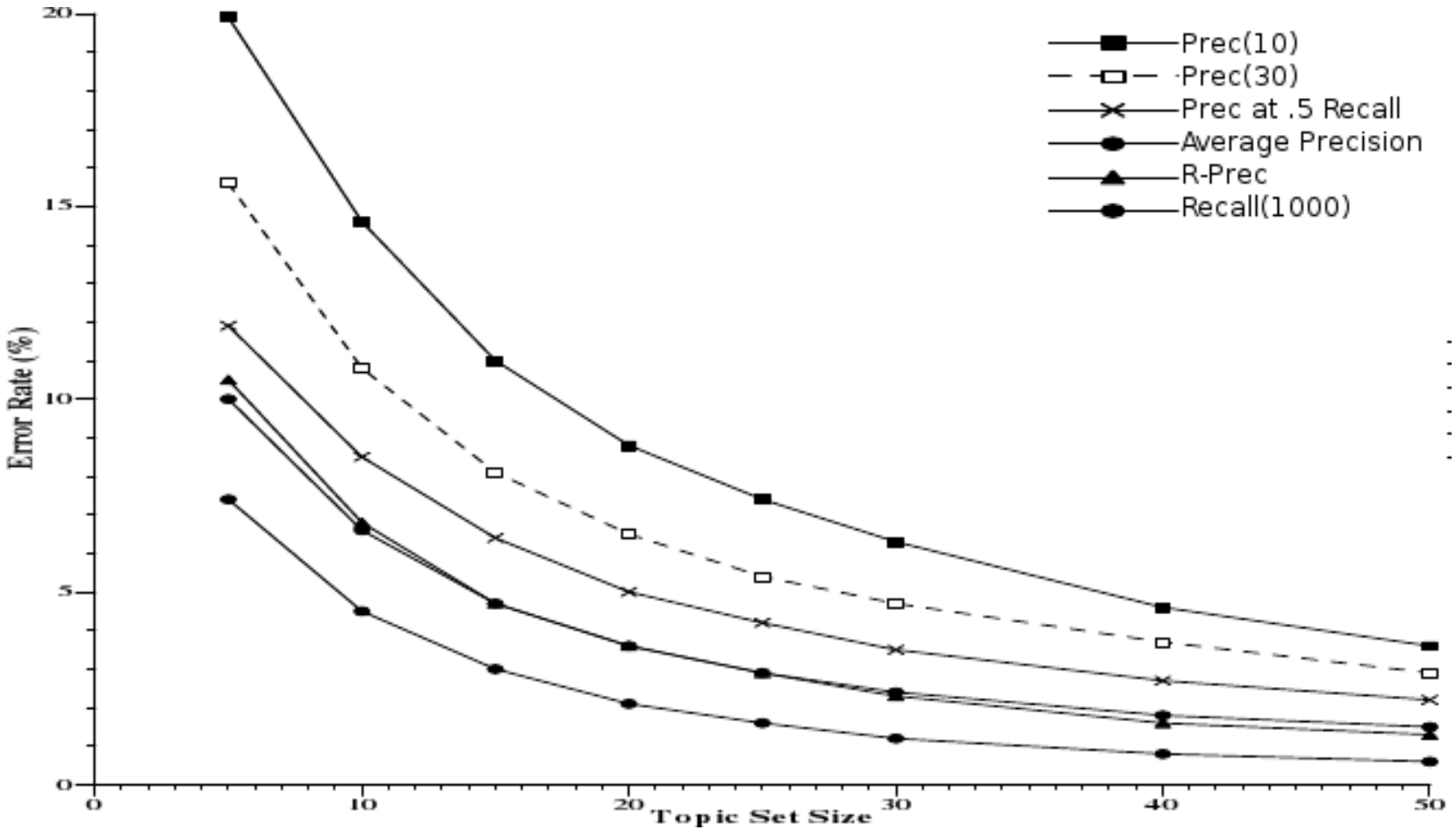

Figure: Average error rate of evaluation measures for varying topic set size

Varying fuzziness values

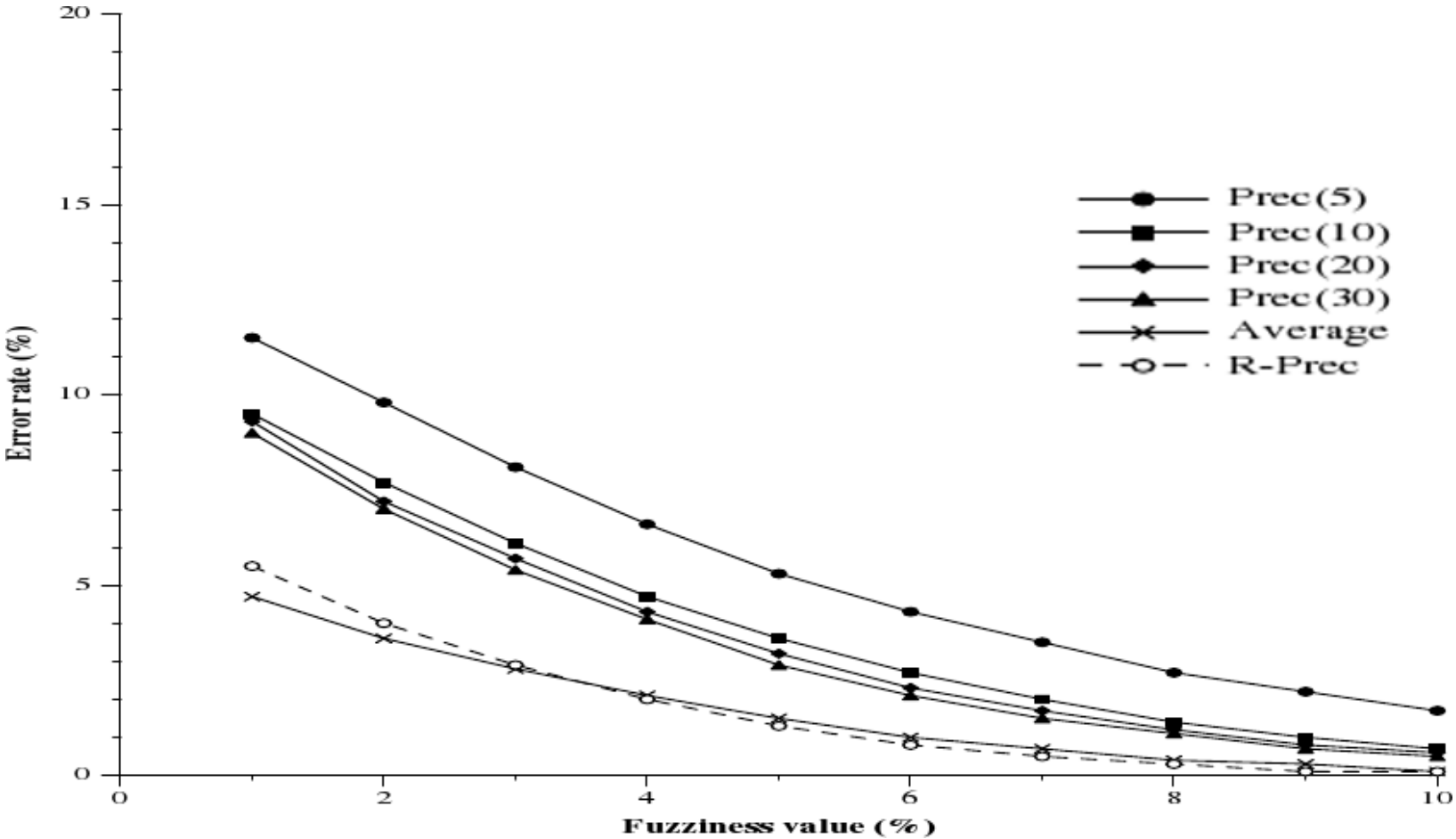

Figure: The effect of fuzziness value on average error rate

Conclusion

- Compare two methods given
  - the number of queries
  - evaluation measure
  - difference threshold

Conclusion

- Some evaluation measures are inherently more stable than others.
  - Precision(1000) is more stable than Avg. Precision.
  - Avg. Precision is more stable than Precision(10).
  - They suggest Avg. Precision.
- Using more queries is more reliable than using fewer queries.
- Requiring a larger threshold between methods increases reliability.
  - But decreases the discrimination between methods.