Evaluating evaluation measure stability by Chris Buckley, Ellen M. Voorhees

(Tunc Gultekin, Eren Golge, Havva Gulay Gurbuz, Ahmet Iscen)

- Comparison of IRSs
  - A is better than B
    - In what sense?
    - What is the measure?
    - Is the difference significant?
- Convincing Measure of Performance
  - Having enough sized test data?
  - Picky evaluation measures?
  - Definition of “fuzziness”* value?
    - * a threshold value that defines if the difference of measures are discriminative enough
- Good Dataset
  - Enough data → stability of the results
  - Query set → well defined vs poor defined
    - Trec 8 Query Track
      - 21 different query sets with 50 different queries
  - Different data comes with different results!
- Measurement Methods
  - Precision CUT_OFF - Prec(c)
  - Precision on 0.5 Recall – Prec 0.5
  - Precision after R - Prec(R)
  - Average Precision for each Relevant Doc. Retrieved – avg.Prec
- Experiment Cycle
  - Pick a query set
    - Pick a fuzziness value
    - Pick measure (etc. Prec-R, avg.Prec ...)
    - Run systems
    - Update comparison matrix
    - Re-pick new query set
    - Do it again
  - Find the error of the matrix
- Structure of Matrix
  - Error* = min(A>B,B>A)/ (A>B + B>A + A==B)
    - * if Error of the comparison matrix >~ 25% than discrimination converges to randomness
- Which measures work?
  - avg.Prec and Prec 0.5 gives less error compared to others (~15-17% vs ~23-25%)
- Error Rate
  - Decreases with …
    - increasing number of topics
    - Increasing fuzziness value (descriptiveness ?)
- Drawbacks !
  - Different result with different test sets
  - Need to have large dataset for stability
- Conclusion
  - Good approach for comparison of IR systems.
  - Need some pre-defined parameters.