Estimating Pool-depth on Per Query Basis

Sukomal Pal, Mandar Mitra, Samaresh Maiti

sukomal_r@isical.ac.in

Information Retrieval Lab, CVPR Unit

Indian Statistical Institute

Kolkata - 700108, India.

http://www.isical.ac.in/~sukomal_r

Background

- Evaluation Test Collection: **Cranfield Method**
  - *corpus*: a set of documents
  - *topics*: a set of information need
  - *qrels*: a set of relevance judgments for each topic
- Exhaustive ground-truth generation *impossible* $\Rightarrow$ POOLing
- examples: TREC, CLEF, NTCIR, INEX, FIRE

Background

- Cranfield Method: Pooling
  - *Biased Sampling* from submissions or runs
  - top-k documents are shortlisted from each of n runs for each topic
  - set-based union of the documents so chosen $\Rightarrow$ POOL
- Exhaustive judgment (relevant or non-relevant) of pool

Motivation

- size of the pool $\sim O(kn)$ per topic
- *e.g.* TREC-8
  - $k = 100, n = 129, qrels = 86,830$ judgments for 50 topics
  - Smaller than corpus-size $= 500,000$
  - BUT cost of evaluation: HIGH
- Cost prohibitive if n & nos. of topics higher
- Solution: low-cost evaluation

Observation: Pooling

Rate of finding new *reldocs* is query-specific !

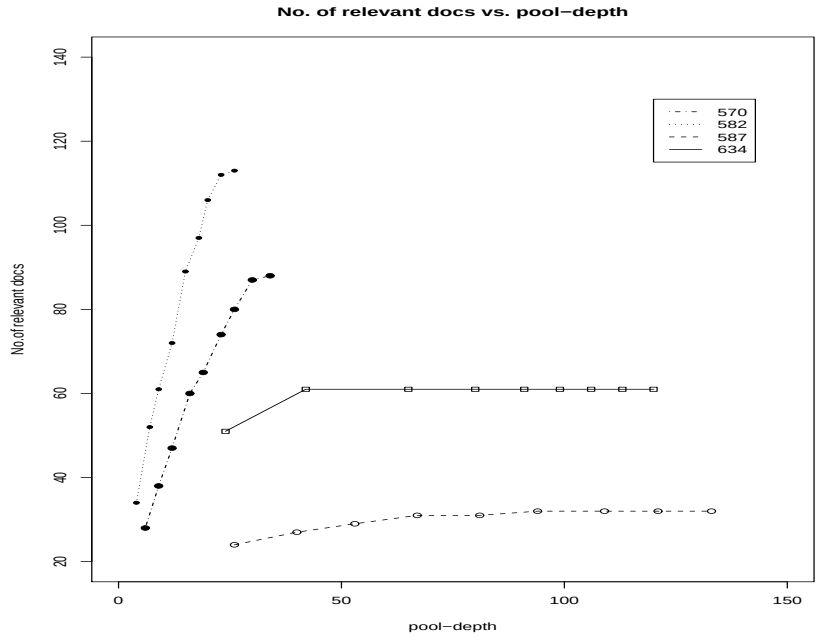


Figure 6: No. of Rel docs vs. Pool-depth.

So is its point of saturation or *critical pool-depth* (k_{cr}) !!

Observation:Pooling

Table 1: Pool saturation at k_{cr}

ad hoc track	**topic-id**	k_{cr}	$nrels$	**pool-size at**	
				k_{cr}	$k = 100$
TREC-7	363	20	16	348	1597
	384	76	51	926	1225
TREC-8	403	14	21	148	1382
	410	47	65	943	2183
NTCIR-5	31	25	32	538	1723
	4	20	10	451	1788

Our Approach: Per Query based Pooling

- *Motivation*: Estimate critical pool-depth (k_{cr})
- Algorithm Overview
  - incrementally build pool from runs starting from $k = 1$ to $k = 100$
  - find *poolsize* and $nrels$ (or $\#reldocs$) at each pool-depth
  - find rate of new $nrels$ at each pool-depth
  - stop if rate drops near zero (no change in $\#reldocs$ for sufficiently long run of pool-depth)

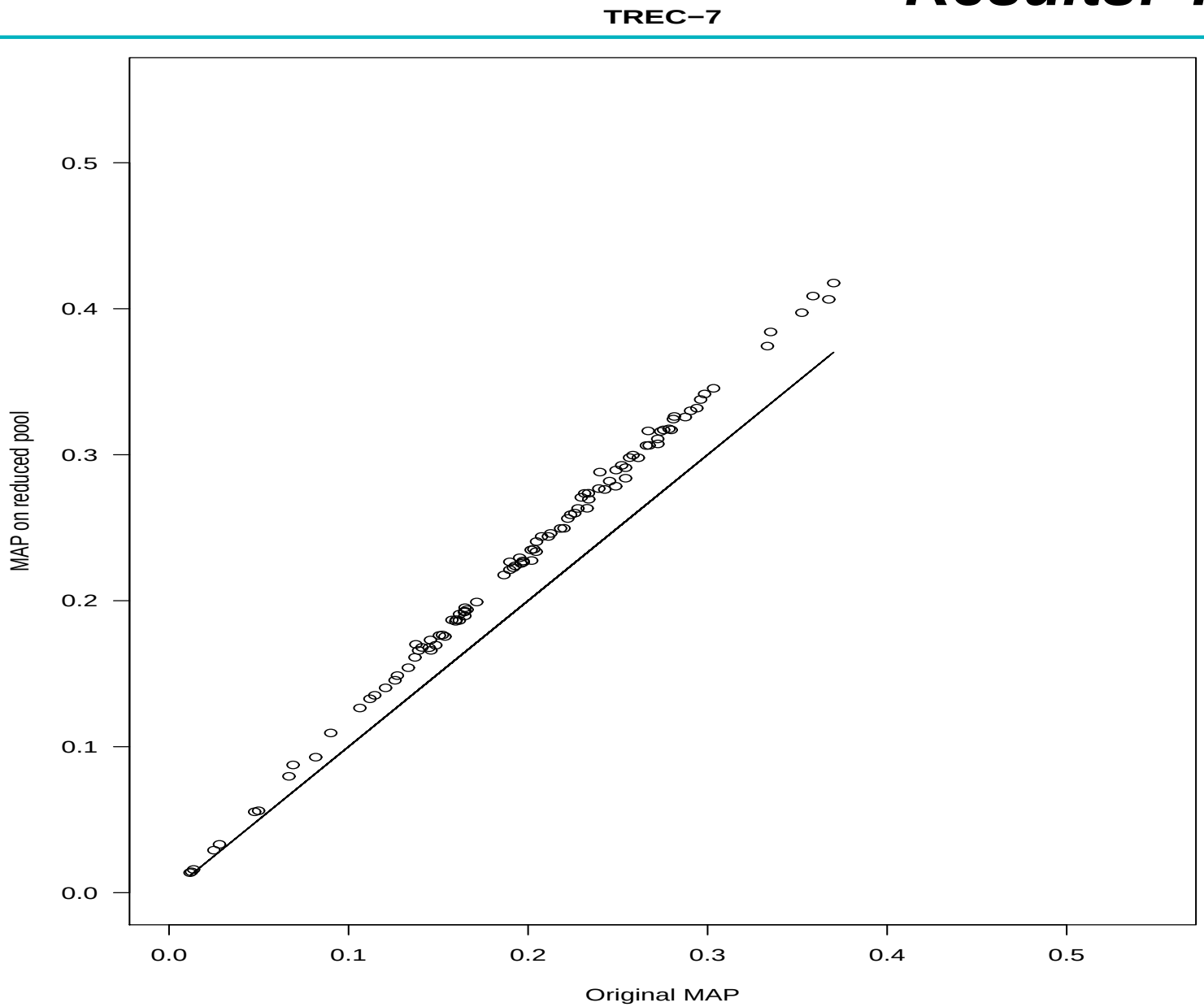
Per Query based Pooling: Algorithm

- as k increases, $nrels$ increases, BUT rate of increase in $nrels$ decreases
- increment in $nrels$ and rate of increment in $nrels$ non-uniform
- 2-stage smoothing
  - smoothing of $nrels$ using window w $(= 6, 8, 10, 12, 14)$
  - smoothing of rate of new $reldocs$ using W $(= 2, 3, 4, 5, 6)$
  - Stop if smoothed rate of 'new' $reldocs <$ threshold t $(= 0.05, 0.10, 0.20, 0.40, 0.80)$ for length l $(= 3, 4, 5, 6)$ of k
- estimate k as k_{cr}

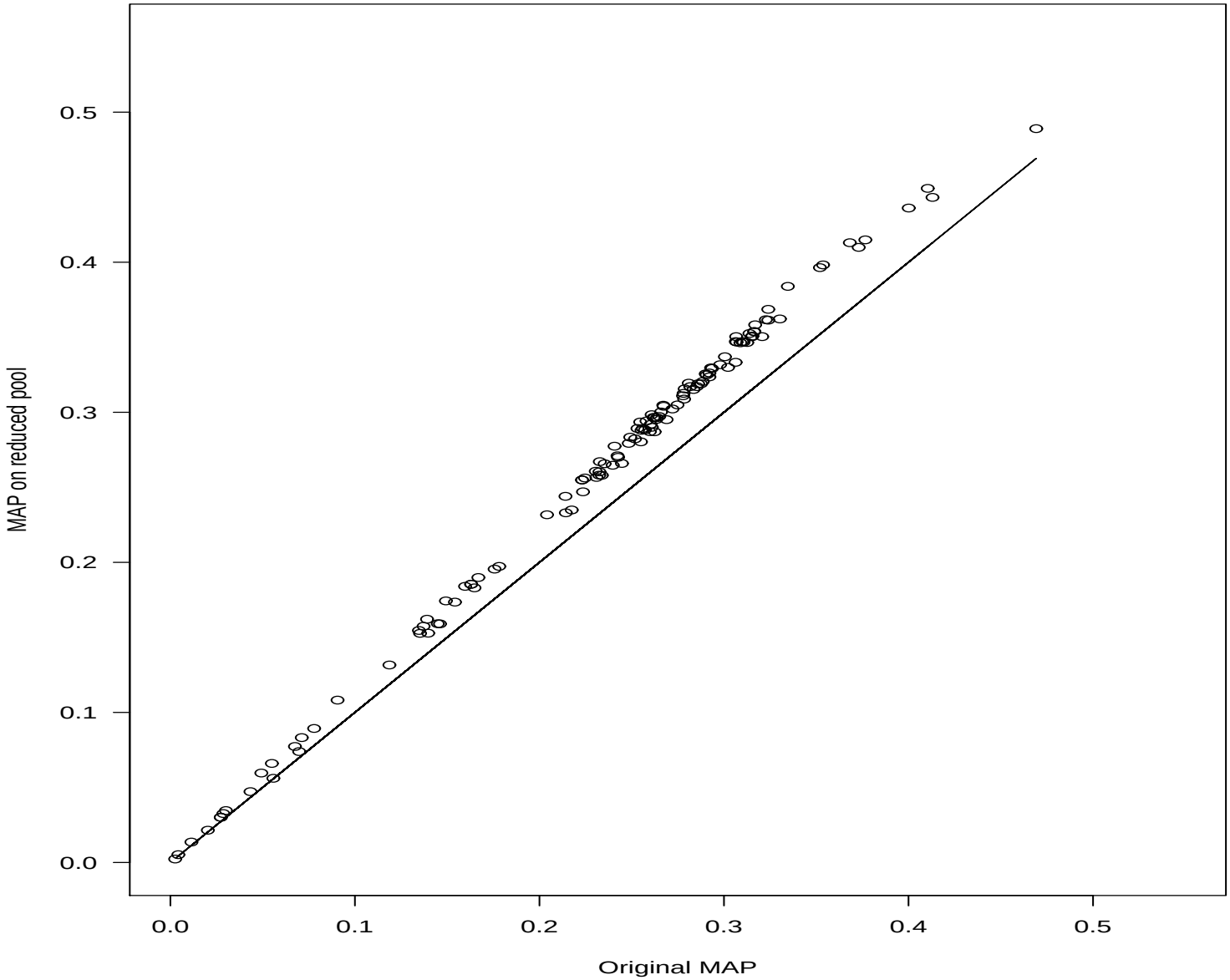
Test data

- *TREC-7:* topics 351-400, 103 runs
- *TREC-8:* topics 401-450, 129 runs
- *NTCIR-5:* topics 1-50, 67 X-E runs (X: J,C,K,E)
- ($5 \times 5 \times 5 \times 4 =$) 500 qrels generated
- All runs evaluated and compared with their original baseline

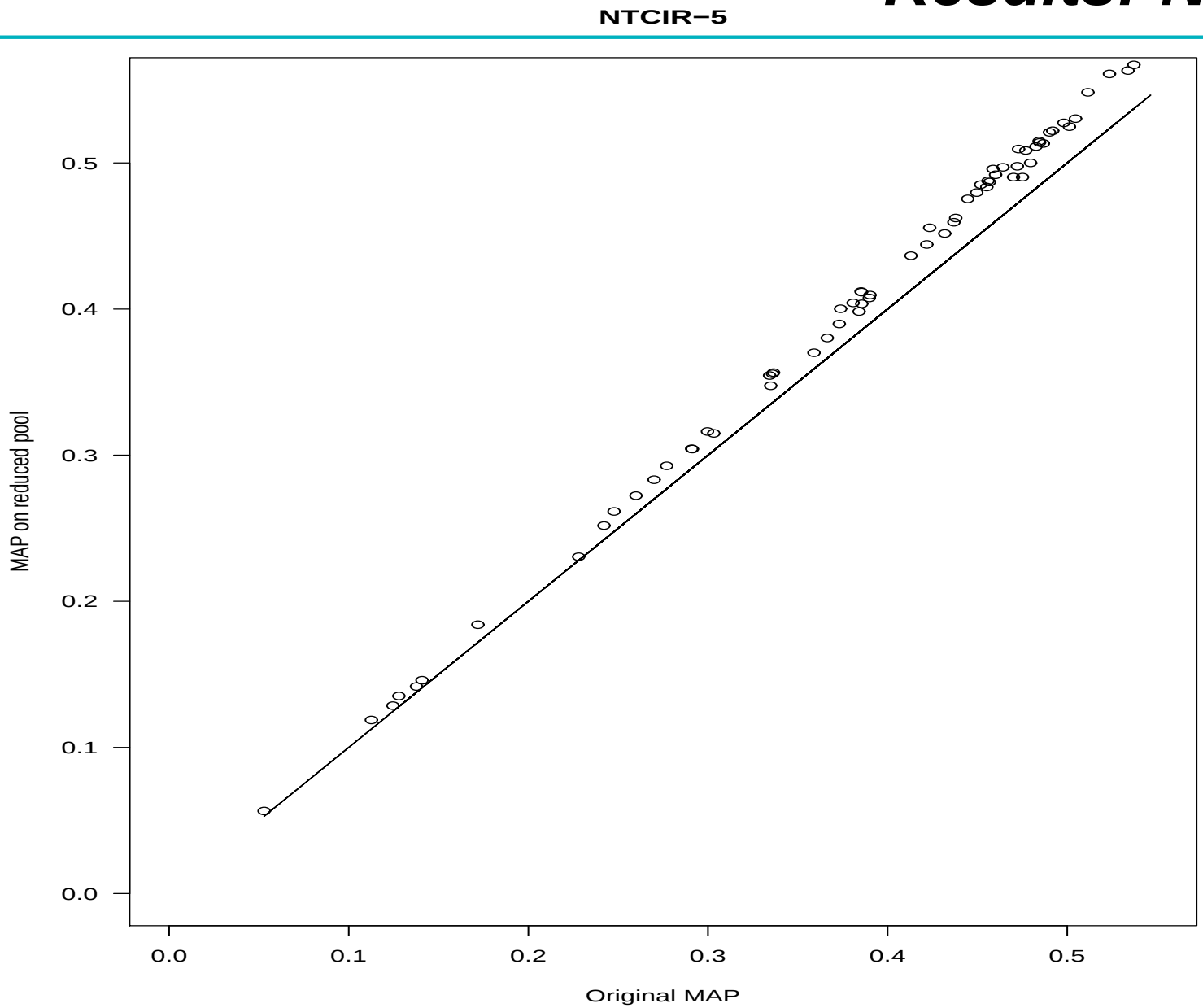
Results: TREC-7



Results: TREC-8



Results: NTCIR-5



Results: Graph-summary

- Graphs show most aggressive stopping criteria (*worst-case* scenario)
- MAPs in close agreement with original MAP(*baseline*)
- New MAPs slightly overestimated
- aggressive stopping $\Rightarrow$ smaller recall-base $\Rightarrow$ increased AP
- MAP difference NOT alarmingly high (lower RMS error)
- relative ranking is more important (higher correlation)

Results: Per Query-based Pooling

Table 2: Guaranteed Performance in reduced pool

track	**Kendall's τ**			**RMS error(ϵ)**		
	τ_{min}	E	R	ϵ_{max}	E	R
TREC-7	0.979	0.381	0.847	0.033	0.379	0.846
TREC-8	0.967	0.368	0.821	0.030	0.369	0.821
NTCIR-5	0.970	0.341	0.850	0.026	0.331	0.846

With respect to original pool
E: fraction of effort
R: ratio of $nrels$.

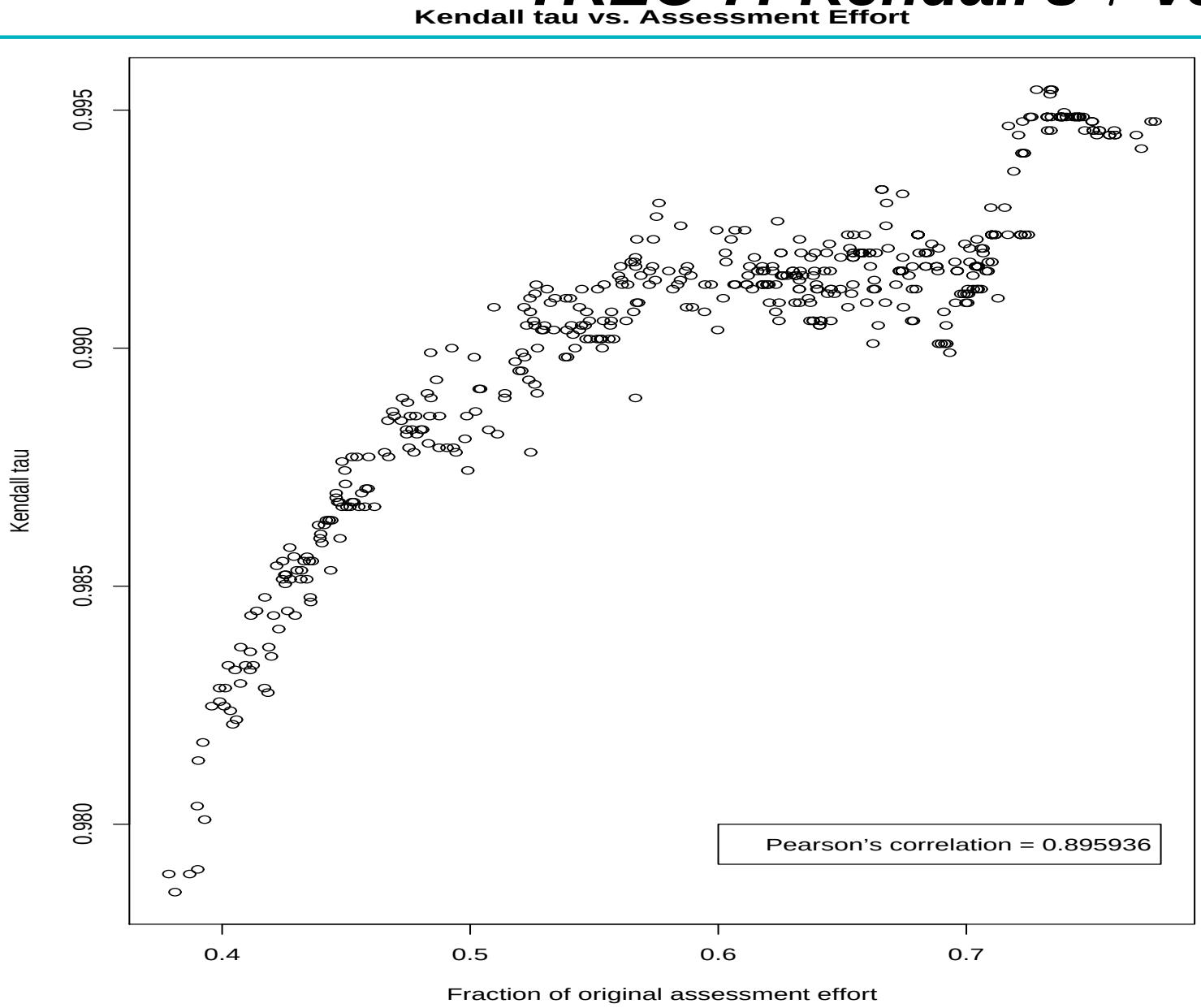
Discussion

- Less than 40% effort
  - identfies more than 80% reldocs
  - produces Kendall's $\tau > 0.96$
  - guarantees less than 3.3% RMS error
- Actual Kendall's τ higher
  - TREC-7: $\tau \in [0.979, 0.996]$
  - TREC-8: $\tau \in [0.967, 0.999]$
  - NTCIR-5: $\tau \in [0.970, 0.996]$
- Actual RMS error lower
  - TREC-7: $\epsilon \in [0.006, 0.033]$
  - TREC-8: $\epsilon \in [0.0009, 0.030]$
  - NTCIR-5: $\epsilon \in [0.002, 0.026]$

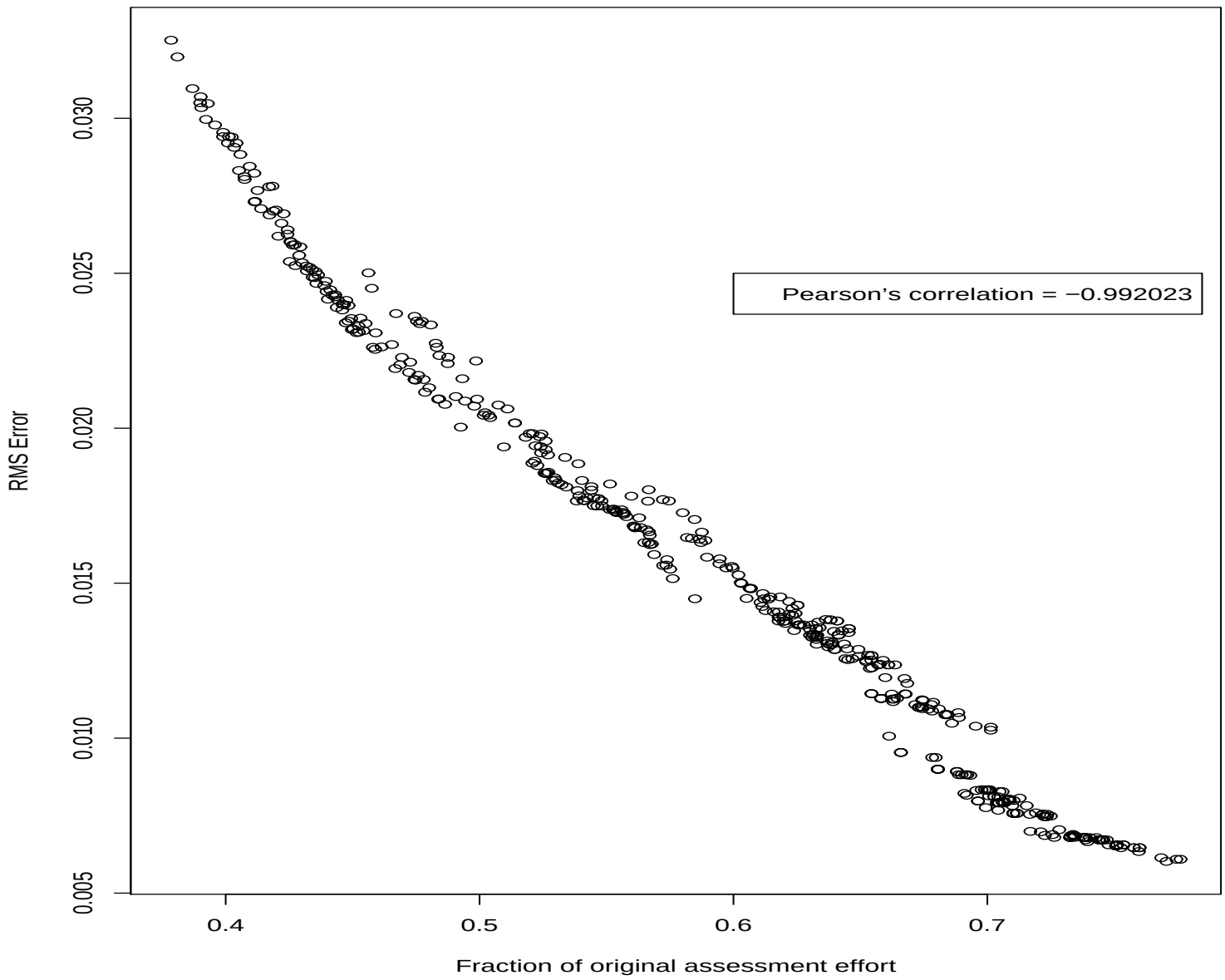
Discussion

- RMS error *inversely varies* as assessment effort
- Rank correlation (τ) *proportional* to assessment effort
- Assessment effort increases with w or W or l
  - if any of w, W or l increases $\Rightarrow$ k_{cr} increases $\Rightarrow$ effort increases
- Assessment effort decreases with increase in t
  - acceptable threshold increases $\Rightarrow$ coarse smoothing $\Rightarrow$lower effort

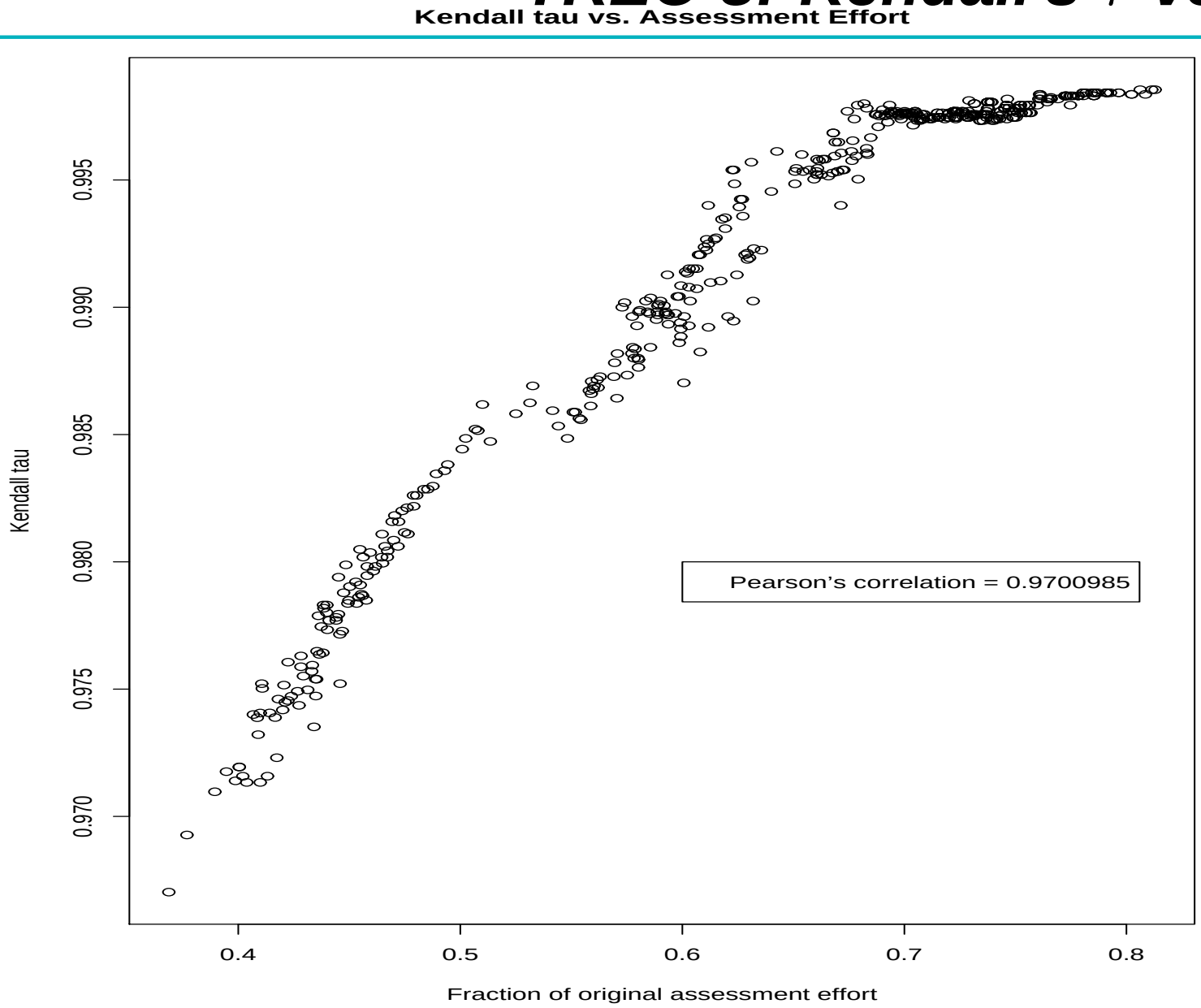
TREC-7: Kendall's τ vs Effort



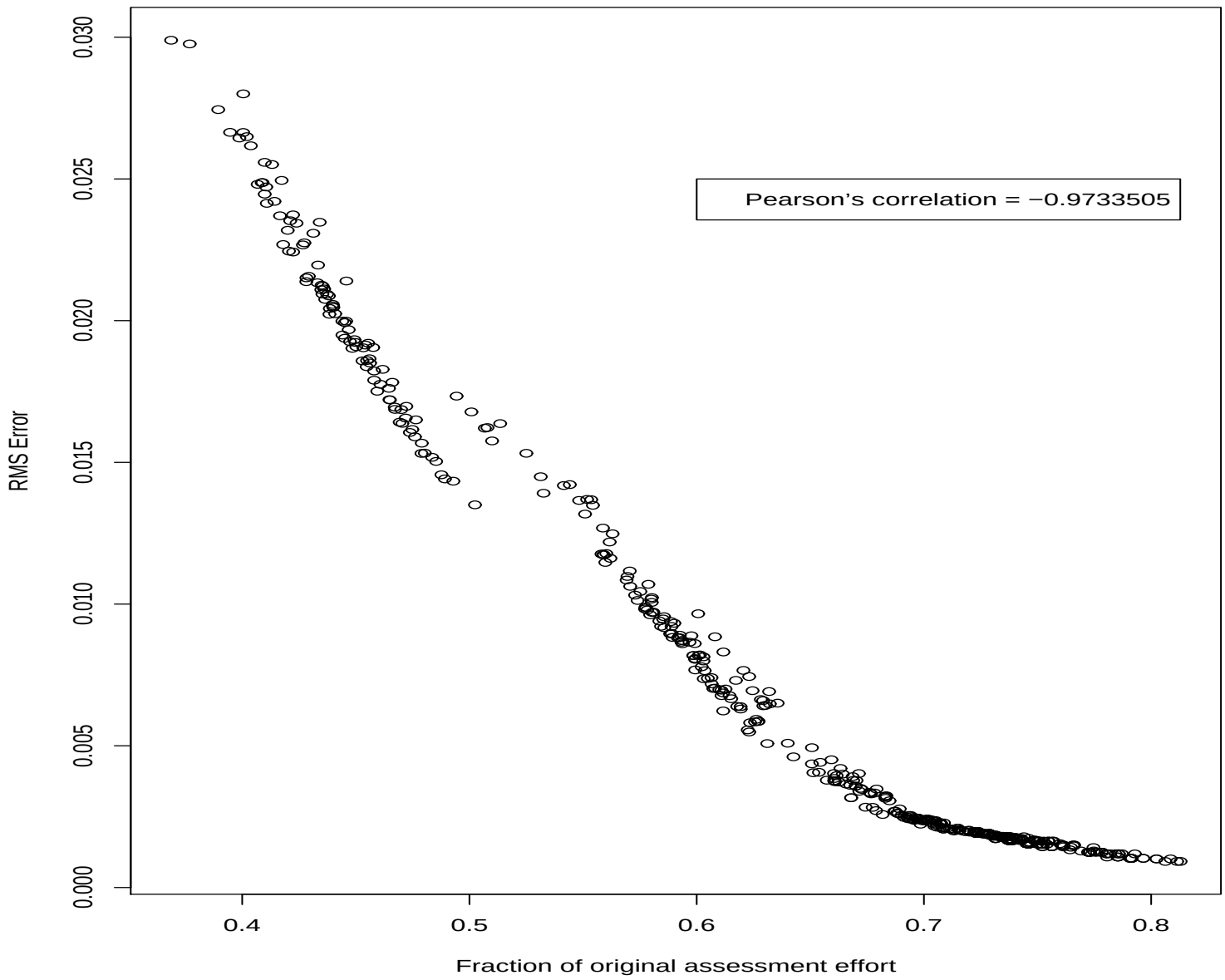
TREC-7: RMS error vs Effort



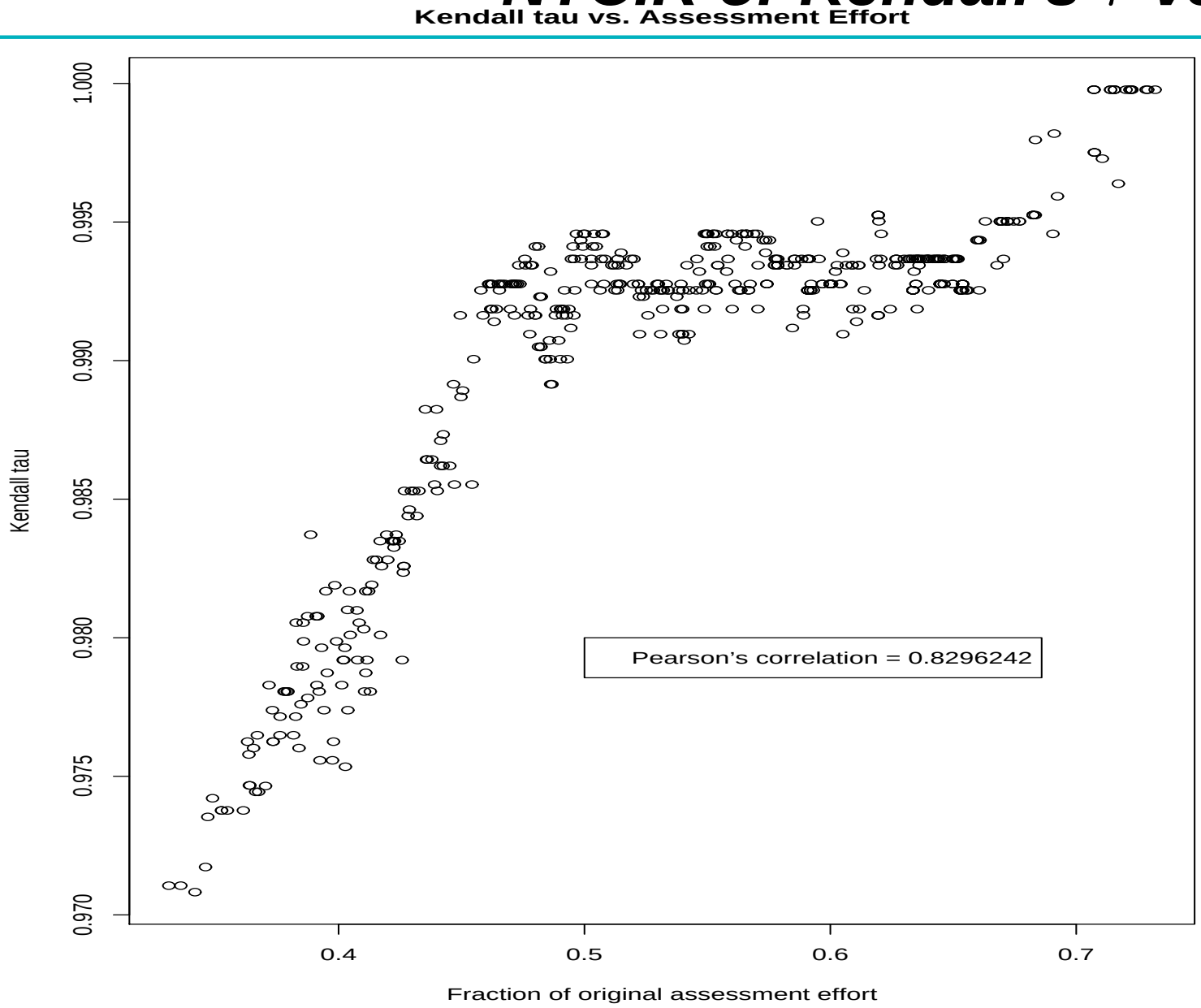
TREC-8: Kendall's τ vs Effort



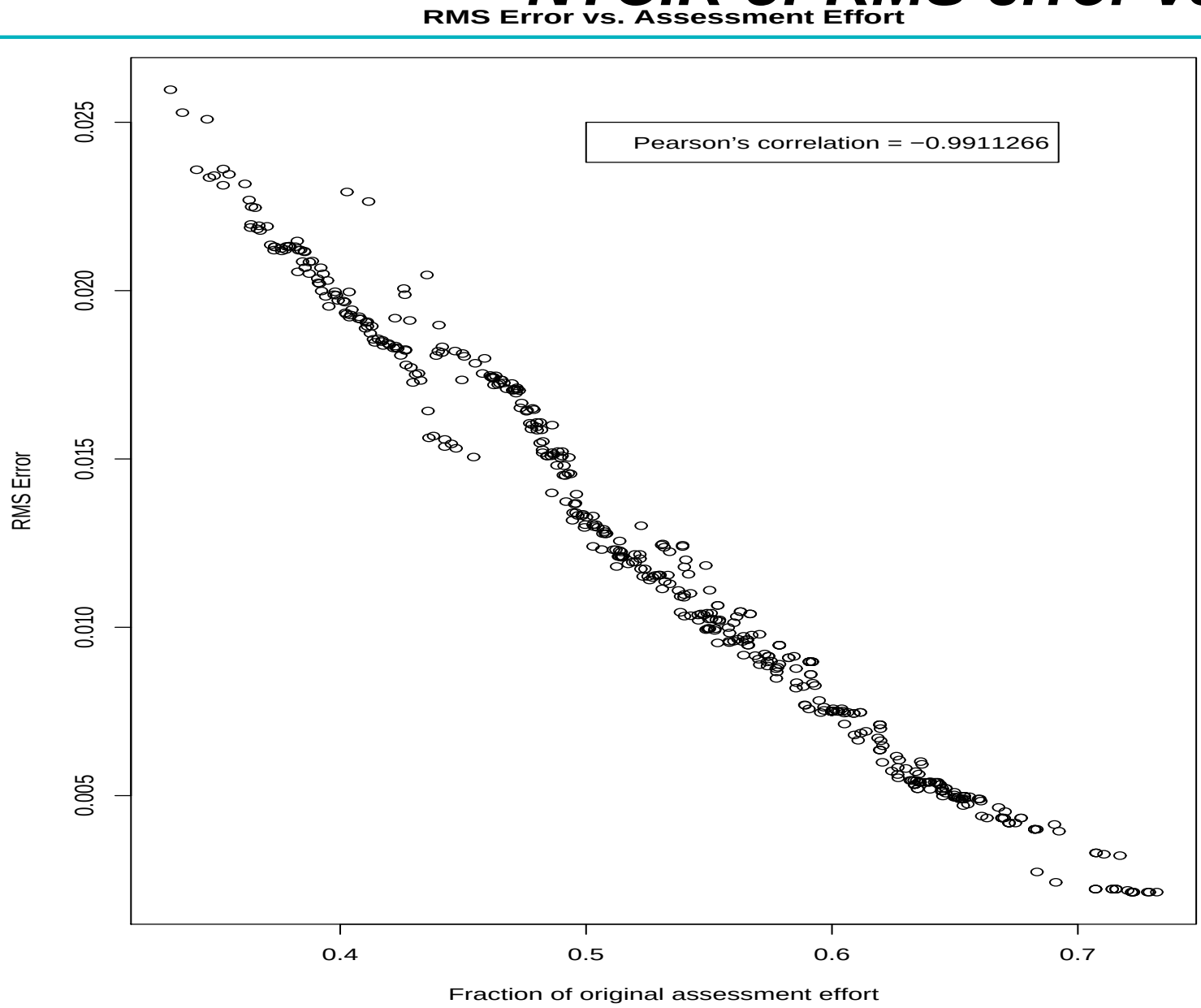
TREC-8: RMS error vs Effort



NTCIR-5: Kendall's τ vs Effort



NTCIR-5: RMS error vs Effort



Conclusion

- Unlike other low-cost evaluation methods, our method is very simple
- For most topics where pool saturates quickly, method pays great dividend
- For topics with high $nrels$, better recall estimates can be achived with high $k(> 100)$
- Tuning 4 parameters (w, W, l, t) gives trade-off betn. cost and reliability
- Reusuability study needs to be done

Acknowledgments

- **Data:** TREC, USA & NTCIR, Japan
- **Work:** Dept. of IT, Govt. of India.
- **Travel:** Google Inc., USA.

!! THANK YOU !!