Magnitude Estimation for Relevance Assessment (& Other Applications)

Kevin Roitero
University of Udine, Italy
roitero.kevin@spes.uniud.it

Sliema, Malta, 14/11/2019

Joint work with

- Shane Culpepper
- Gianluca Demartini
- Eddy Maddalena
- Stefano Mizzaro
- Mark Sanderson
- Falk Scholer
- Andrew Turpin

Outline

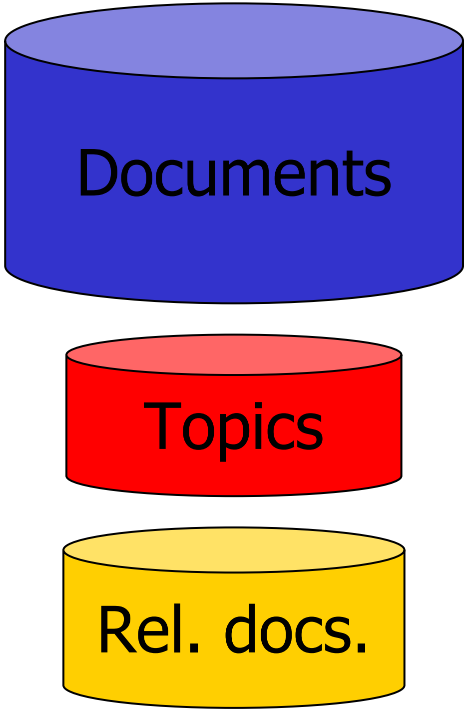
- Intro
  - 1. IR evaluation & Relevance assessment
  - 2. Crowdsourcing
  - 3. Relevance scales
- Scale 1: Magnitude Estimation
  - Experiments, Results…
- Scale 2: 100 values
  - Experiments, Results…
- other applications and ongoing work
- Conclusions

IR

- Information Retrieval
- You know what it is right?
- Google.

Benchmark-based IR evaluation

- Cranfield, TREC, NTCIR, INEX, CLEF, FIRE, …
- “Test collection”
  - Documents
  - Topics (queries)
  - System output as ranked docs
  - Relevance judgments (by humans)



Relevance and IR Evaluation

- Relevance is a fundamental concept in IR
- Plays a key role in the evaluation of IR systems
- But is complex, and difficult to operationalize
  - Judgments often based only on “topical” relevance
    - Novelty? Context? Authority?...
  - Choice of scale
    - Binary? Ordinal? (3? 4? 5? 7? 21?)…

2. Crowdsourcing

- "Outsource to the crowd"
- "taking a task traditionally performed by an employee or contractor, and outsourcing it to an undefined, generally large group of people or community in the form of an open call" [http://en.wikipedia.org/wiki/Crowdsourcing]

3. Relevance scales

- How to express relevance?
- Traditional solution: **Binary** relevance

- **4-level** ordinal scale [Sormunen, 2002]
  - 3: Highly relevant (H)
  - 2: Relevant (R)
  - 1: Marginally relevant (M)
  - 0: Not relevant (N)

Outline

- Intro
  - 1. IR evaluation & Relevance assessment
  - 2. Crowdsourcing
  - 3. Relevance scales
- Scale 1: Magnitude Estimation
  - Experiments, Results…
- Scale 2: 100 values
  - Experiments, Results…
- other applications and ongoing work
- Conclusions

Magnitude Estimation

- A psychophysical scaling technique for measuring perception
- Stimuli at different levels of intensity are presented to an observer

- The intensity of each stimulus is rated by the assignment of a number, depending on the perceived intensity
- Developed by Stanley Stevens at Harvard in 1950s

Magnitude Estimation…

- Unlimited (≠ “category scale with many categories”)
  - either $]-\infty, +\infty[$ or
  - **$]0, +\infty[$** (we used this one)
- Leads to ratio scale: ratios of the assigned numbers is what’s important
- The granularity of the scale is chosen by the judge, and not constrained by predetermined levels
- Judges can not run out of categories
- used for (physical) stimuli: medicine, law, sociology, linguistics ...

Research Questions

1. Is ME OK for gathering relevance judgments?
2. Do ME and Crowdsourcing “mix well”?
3. What is the effect on system ordering?
4. Can ME scores provide insight into
   - user perceptions of relevance?
   - individual gain profiles?

The Crowd

- User study carried out using CrowdFlower
  (was Figure Eight)
  (now closed for academics)
  (...)
- Each *task unit* required completing a practice task, and assessing 8 documents (on 1 topic)
- Participants paid US$ 0.2 per task
- “Great” CF workers were selected

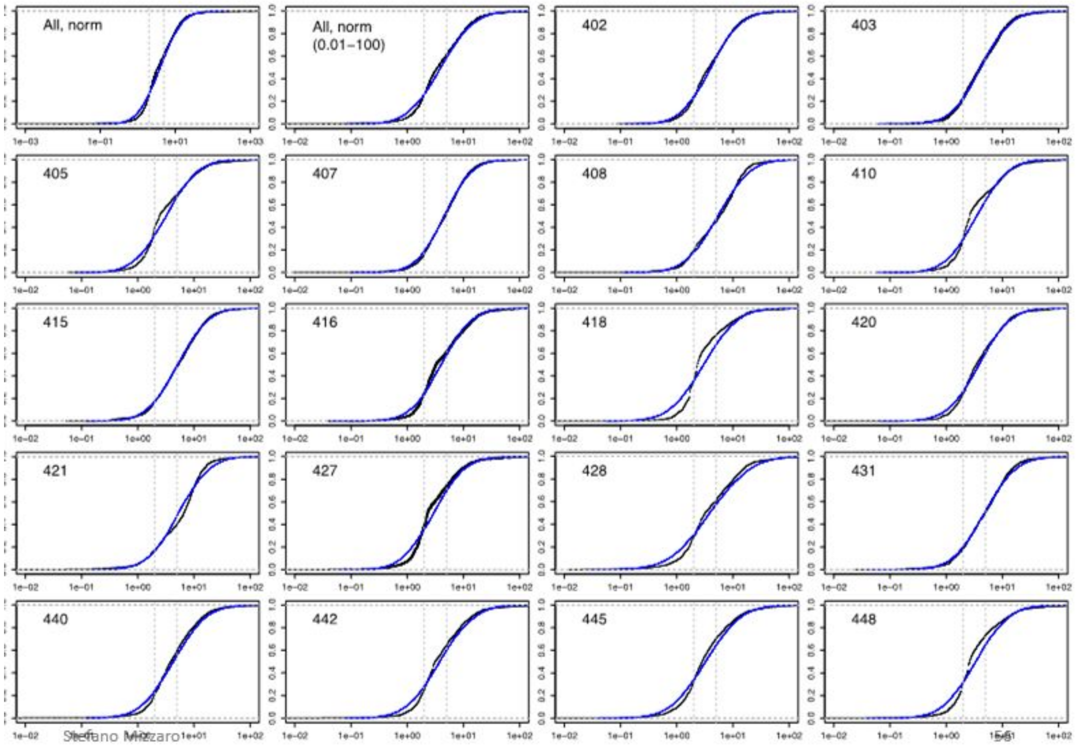
Documents, Topics and Relevance

- Top-10 documents returned by TREC-8 systems
- 18 topics with **binary** judgments…
- … and with judgments on a **4-level ordinal** scale [Sormunen, 2002]
- 10 ME scores gathered for each of the 4,269 topic-document pairs
- Total units: 7,059, ~50k judgments

Main Task

- instructions displayed, including an explanation of ME process
  - “use a ratio scale”
  - “avoid order bias”
- Practice task (3 lines of different length in ascending order)
- Quality checks
  - comprehension test
  - two gold questions
  - repeat 8 times
  - Time spent on each document

Individual scores (normalized)

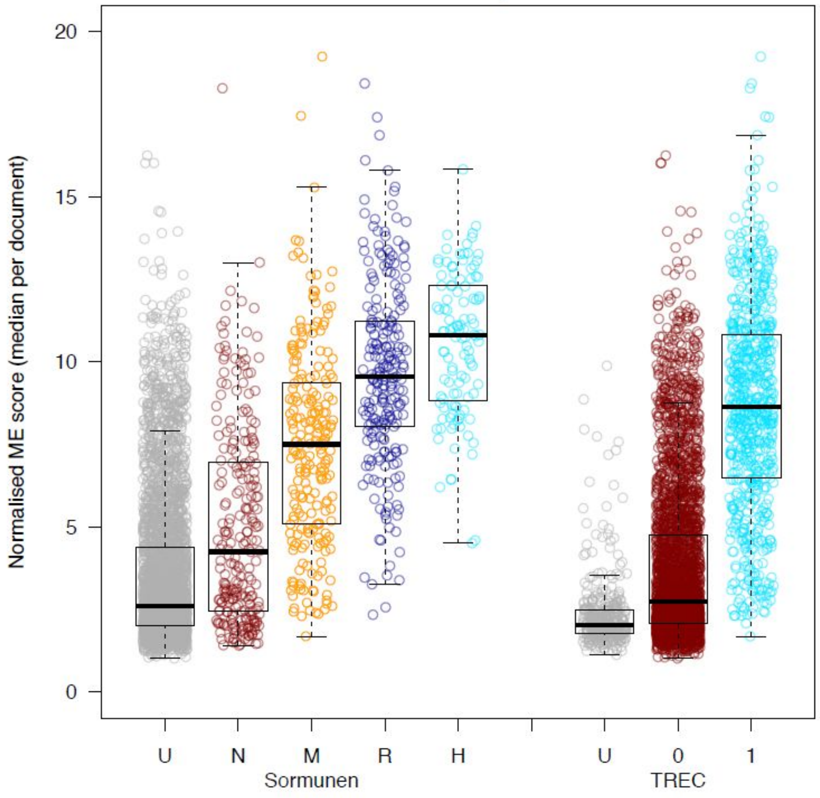


Stefano Mizzaro

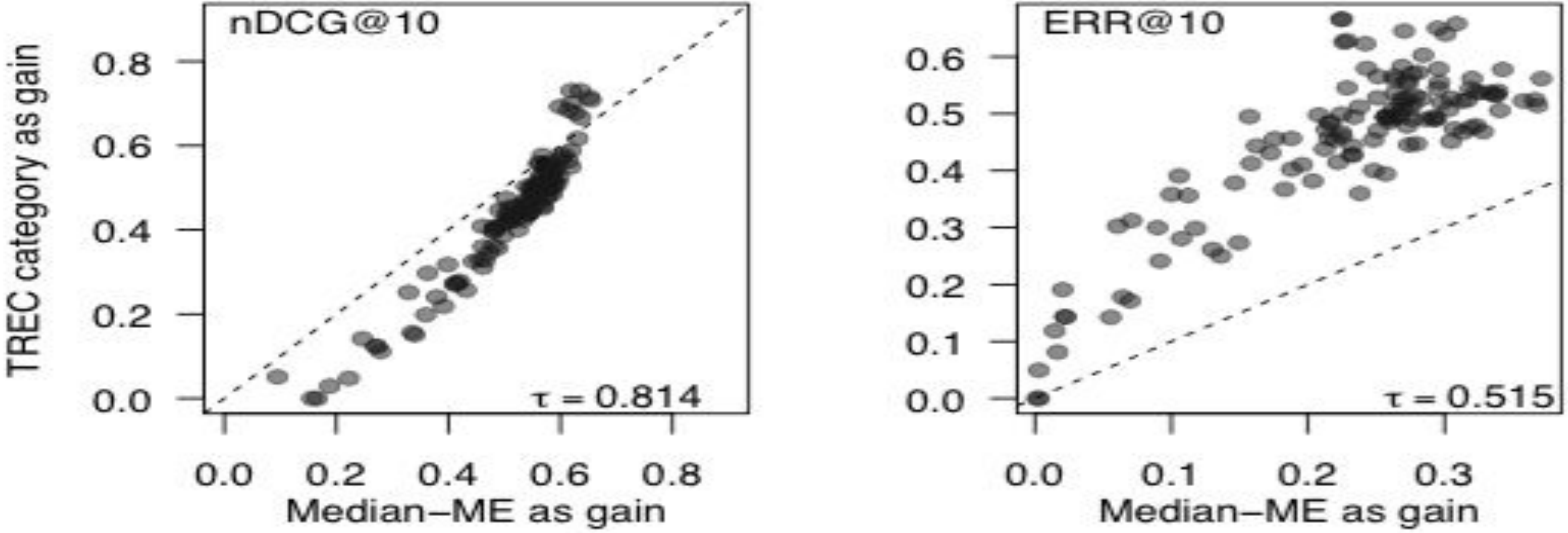
Aggregation

- Aggregation with Median function
- Summary:
  - Workers scores
  - Geometric averaging normalization
  - 10 redundant normalized scores
  - Median

Consistency of ME and Ordinal / Binary Relevance

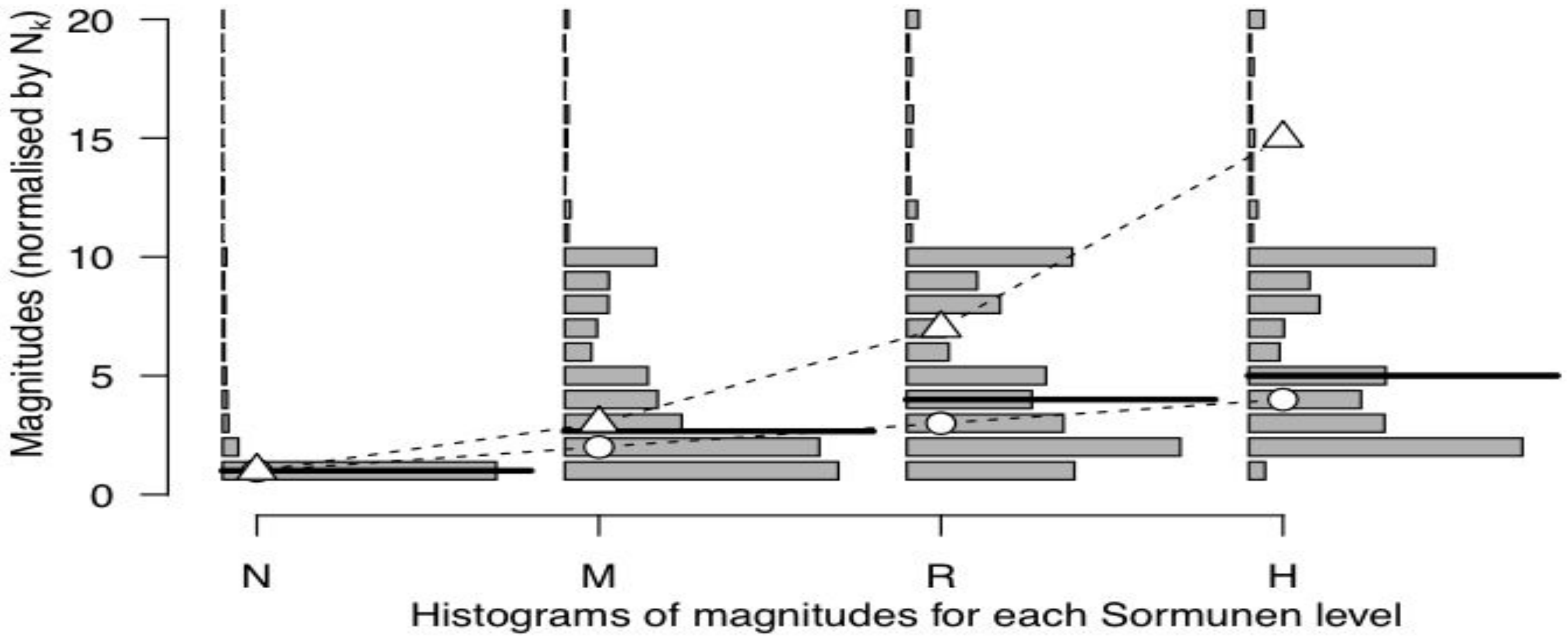


Comparative System Rankings: Binary



Overlap in top set (runs that are statistically indistinguishable from the best run): nDCG: 44%, ERR: 76%

Individual Relevance Perception: Gain Profiles



Research Questions

1. Is ME OK for gathering relevance judgments?
2. Do ME and Crowdsourcing "mix well"?
3. What is the effect on system ordering?
4. Can ME scores provide insight into
   - user perceptions of relevance?
   - individual gain profiles?

Yes!

Yes!

Yes!

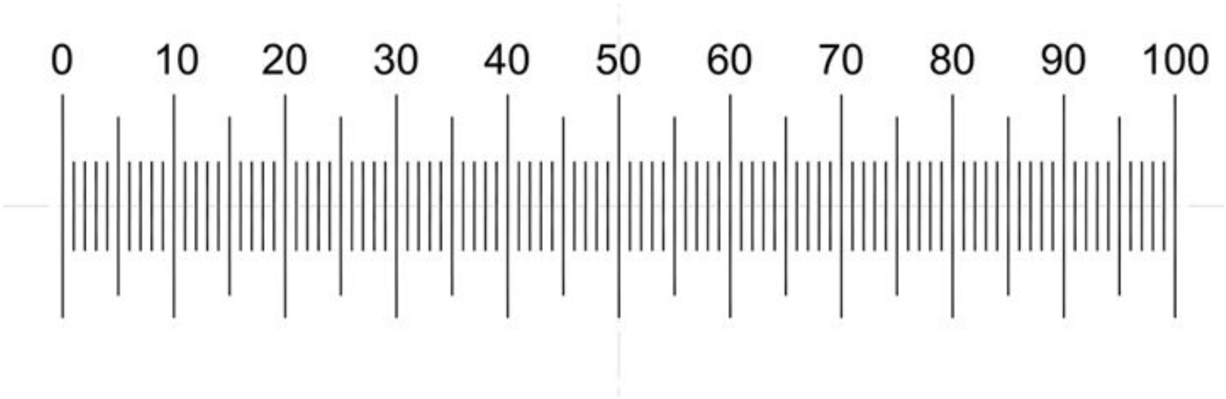
Yes!

Outline

- Intro
  - 1. IR evaluation & Relevance assessment
  - 2. Crowdsourcing
  - 3. Relevance scales
- Scale 1: Magnitude Estimation
  - Experiments, Results…
- Scale 2: 100 values
  - Experiments, Results…
- other applications and ongoing work
- Conclusions

Maybe ME is… too much?

- Let's try with just 100
- …actually 101…
- …We call it S100 anyway



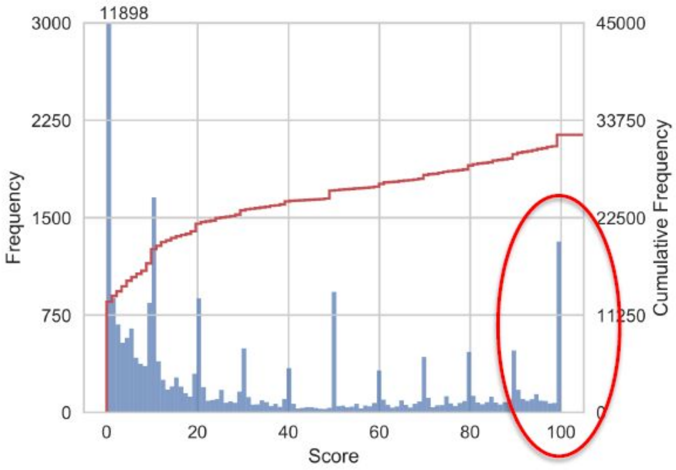
ME vs S100, in theory

- Pro ME
  - Ratio scale
  - New values always available
- Pro S100
  - No normalization issues
  - More familiar / similar to usual approaches (e.g., 5 stars)

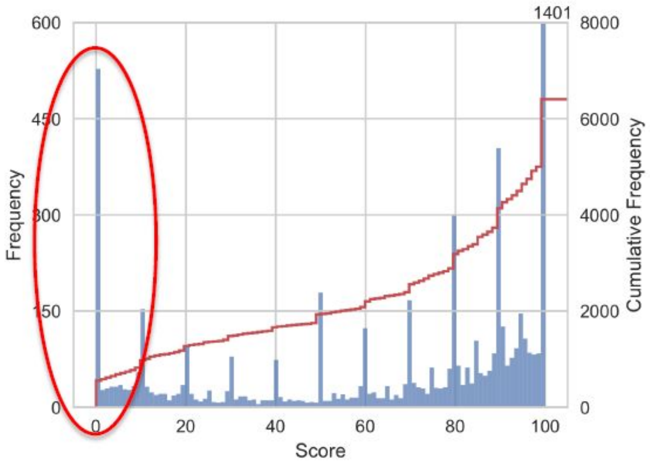
Experiments

- Experimental design: Identical to ME
  - Again on CrowdFlower/Figure8/Whatever
  - Same topics
  - Same documents
  - Same order
  - Same user interface (with minor adjustments)
  - … Some results

Individual scores



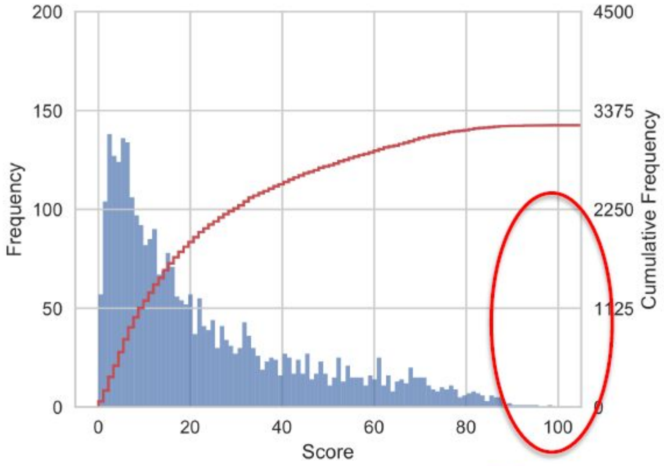
TREC non rel 👎



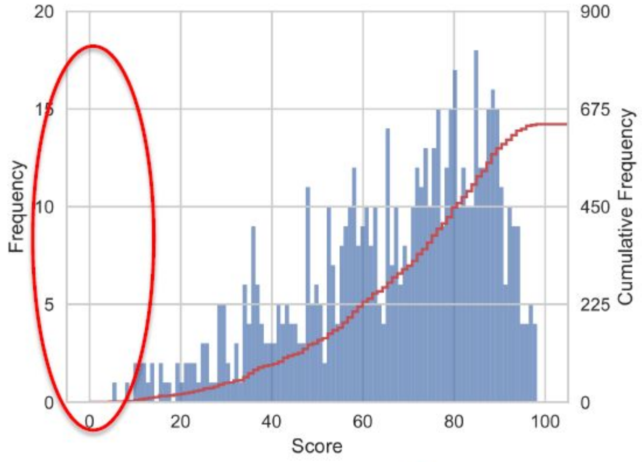
TREC rel 👍

- "Wrong" scores…

Aggregated scores (mean)



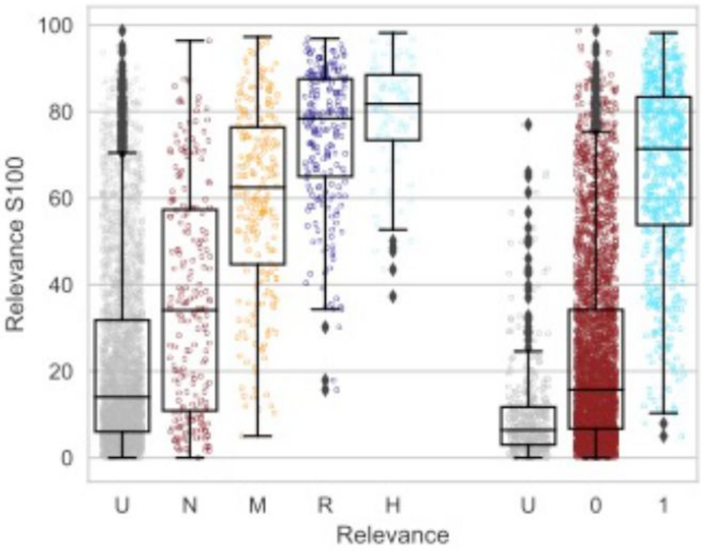
TREC non rel 👎

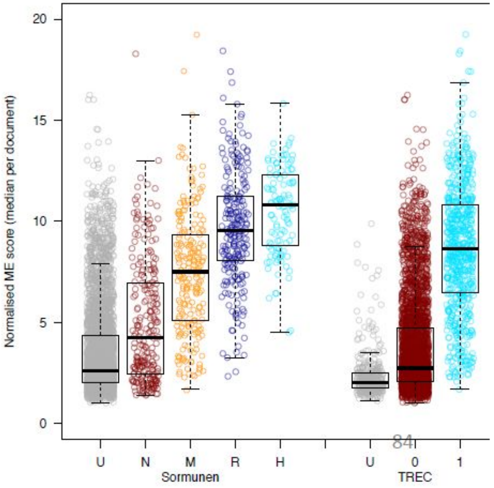


TREC rel 👍

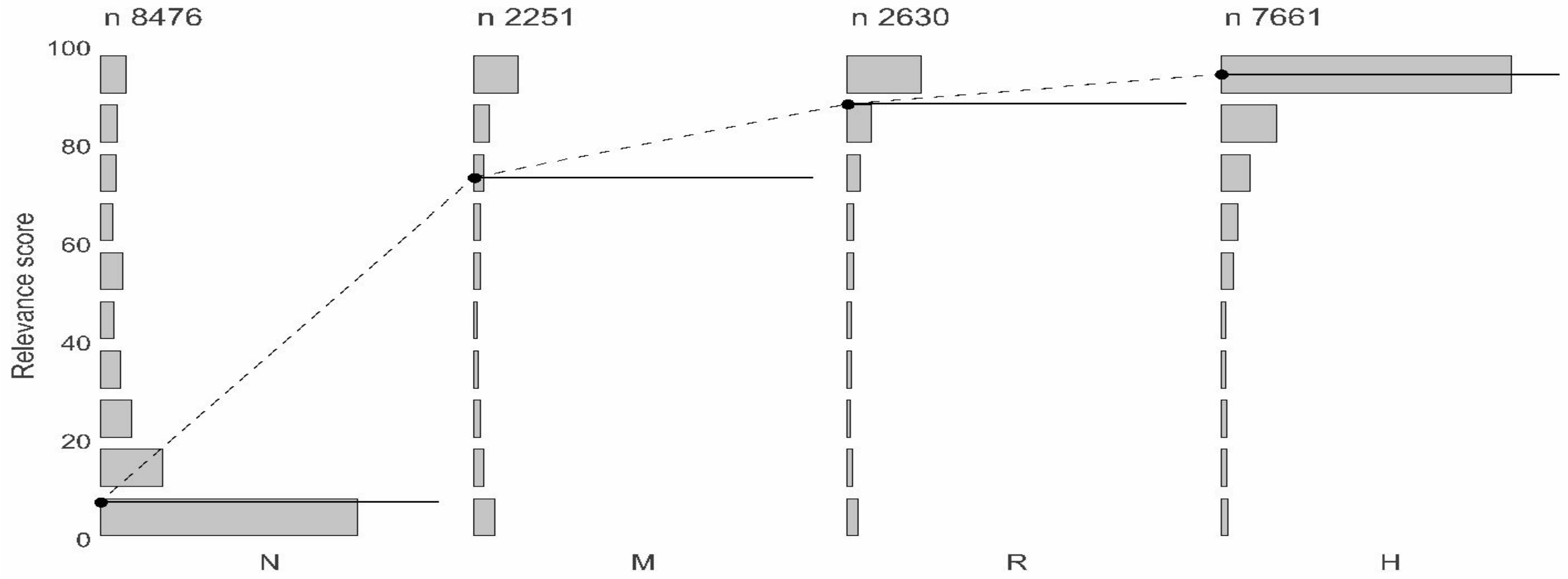
- No more "wrong" scores...

Consistency of S100 and Ordinal / Binary Relevance

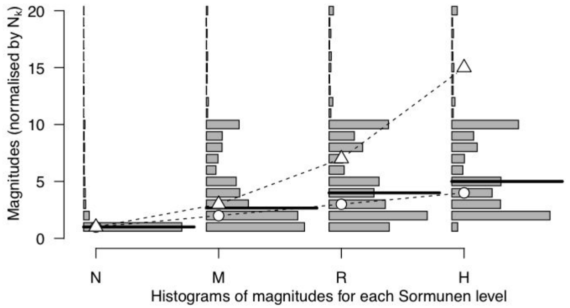




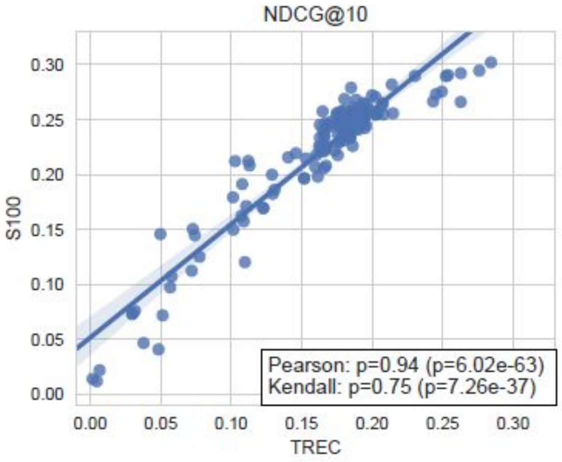
Gain Profiles

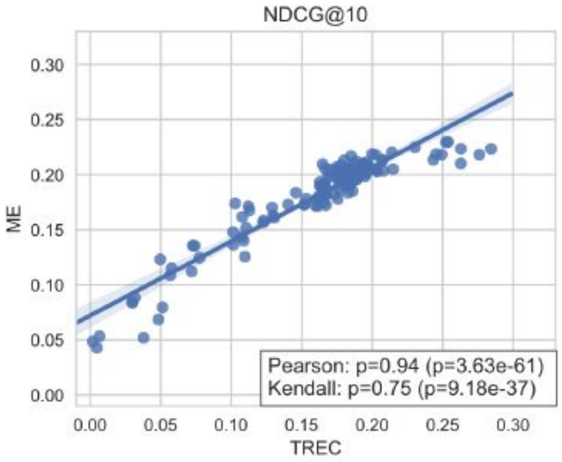


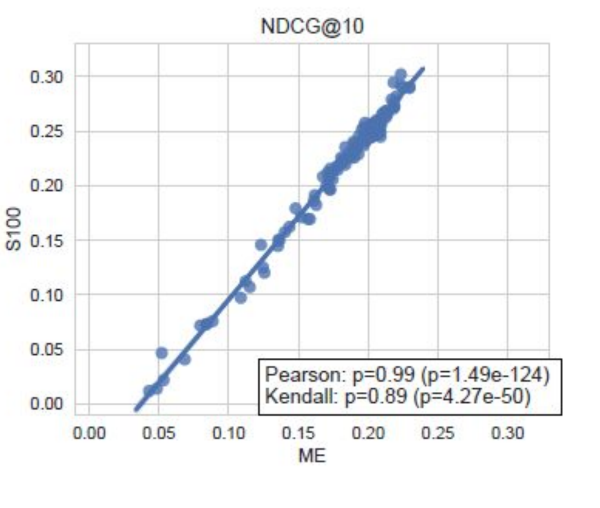
- S100: sublinear
- ME: superlinear



Effect on system ranking

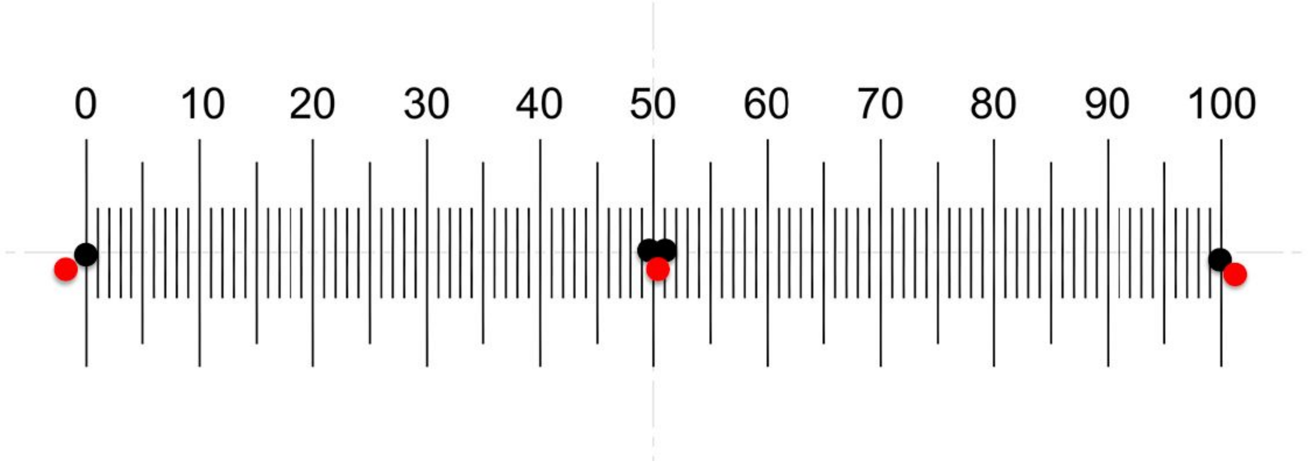






S100 running out of values?

- Scale boundaries
- Discrete vs. Continuous scale
- Rather limited effects



Conclusions / ME]0, +∞[

- ME is a suitable to gather relevance judgments
- ME can be used in a Crowdsourcing setting – at least with some precautions
  - Quality checks, normalization & aggregation, …
- Key advantage: can capture fine-grained variation in human perceptions of relevance
- System orderings vary substantially depending on relevance scale (ME/Binary: $\tau = 0.81$; ME/Ordinal: $\tau = 0.53$)
- Gain profiles
  - Linear seems to match "median" user
  - But lots of variability, one profile seems too simplistic

Conclusions / S100 [0, 100]

- S100 has many of the advantages of ME
- S100 is better w.r.t.:
  - Agreement with TREC (not shown)
  - Familiarity
  - More robust to fewer data (not shown)
- Disadvantages look only theoretical
  - "running out of values" rarely an issue in practice
- S100 looks a good compromise

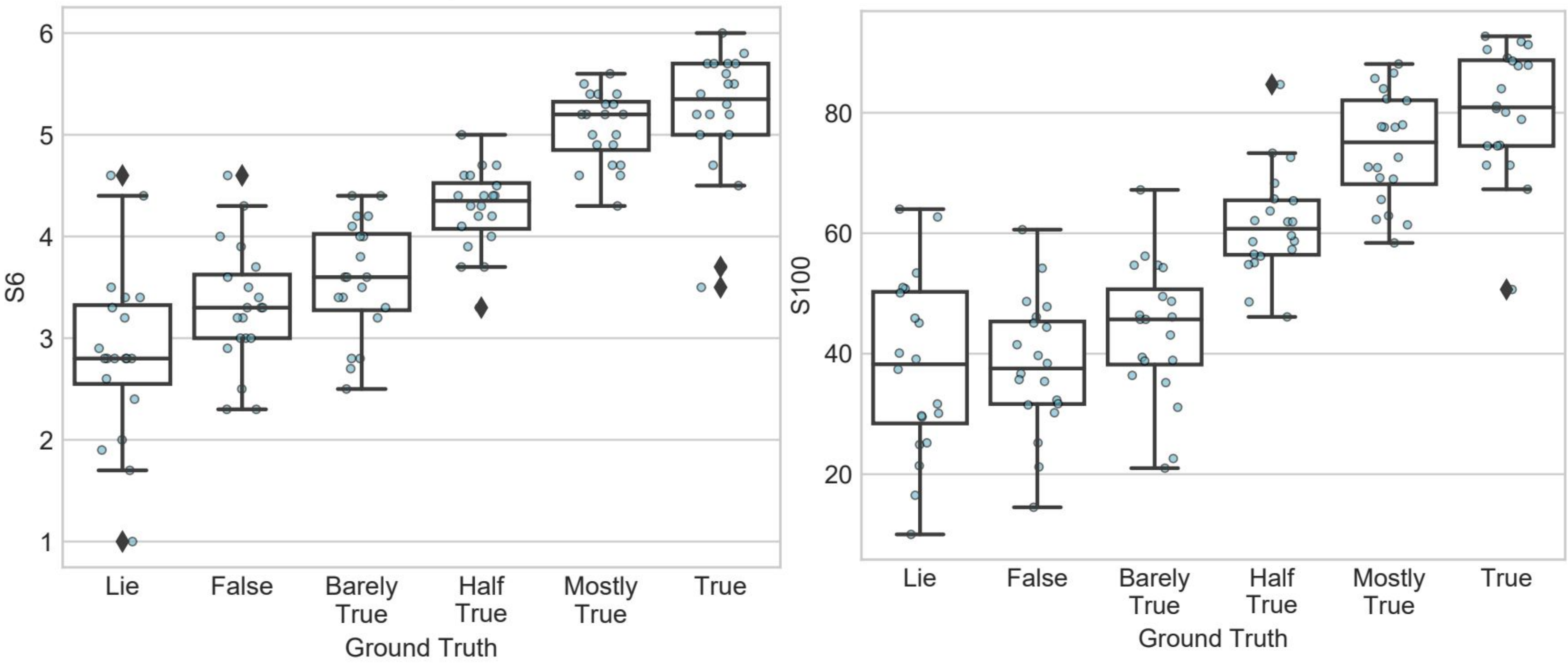
Outline

- Intro
  - 1. IR evaluation & Relevance assessment
  - 2. Crowdsourcing
  - 3. Relevance scales
- Scale 1: Magnitude Estimation
  - Experiments, Results…
- Scale 2: 100 values
  - Experiments, Results…
- other applications and ongoing work
- Conclusions

(not so) Future Work

- Further analyses
- S2 and S4 from the crowd
- S10 as well
- Application to other domains: Fake News Detection

Sneak Preview (Fake News)



References

- Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. **The Benefits of Magnitude Estimation Relevance Assessments for Information Retrieval Evaluation.** ACM SIGIR 2015.
- Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. **On Crowdsourcing Relevance Magnitudes for Information Retrieval** Evaluation. ACM TOIS 2017.
- Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. **On Fine-Grained Relevance Scales**. ACM SIGIR 2018 .
- Kevin Roitero, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. **How Many Truth Levels? Six? One Hundred? Even more? Validating Truthfulness of Statements via Crowdsourcing.** RDSM 2018.
- (and some work on scales transformation)
  - Lei Han, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro and Gianluca Demartini. **On Transforming Relevance Scales.** CIKM2019
- (... and an ECIR Submission)
- (... and a journal paper in preparation)

Thanks

- Co-authors (especially Stefano for… some slides!)
- Reviewers
- You!