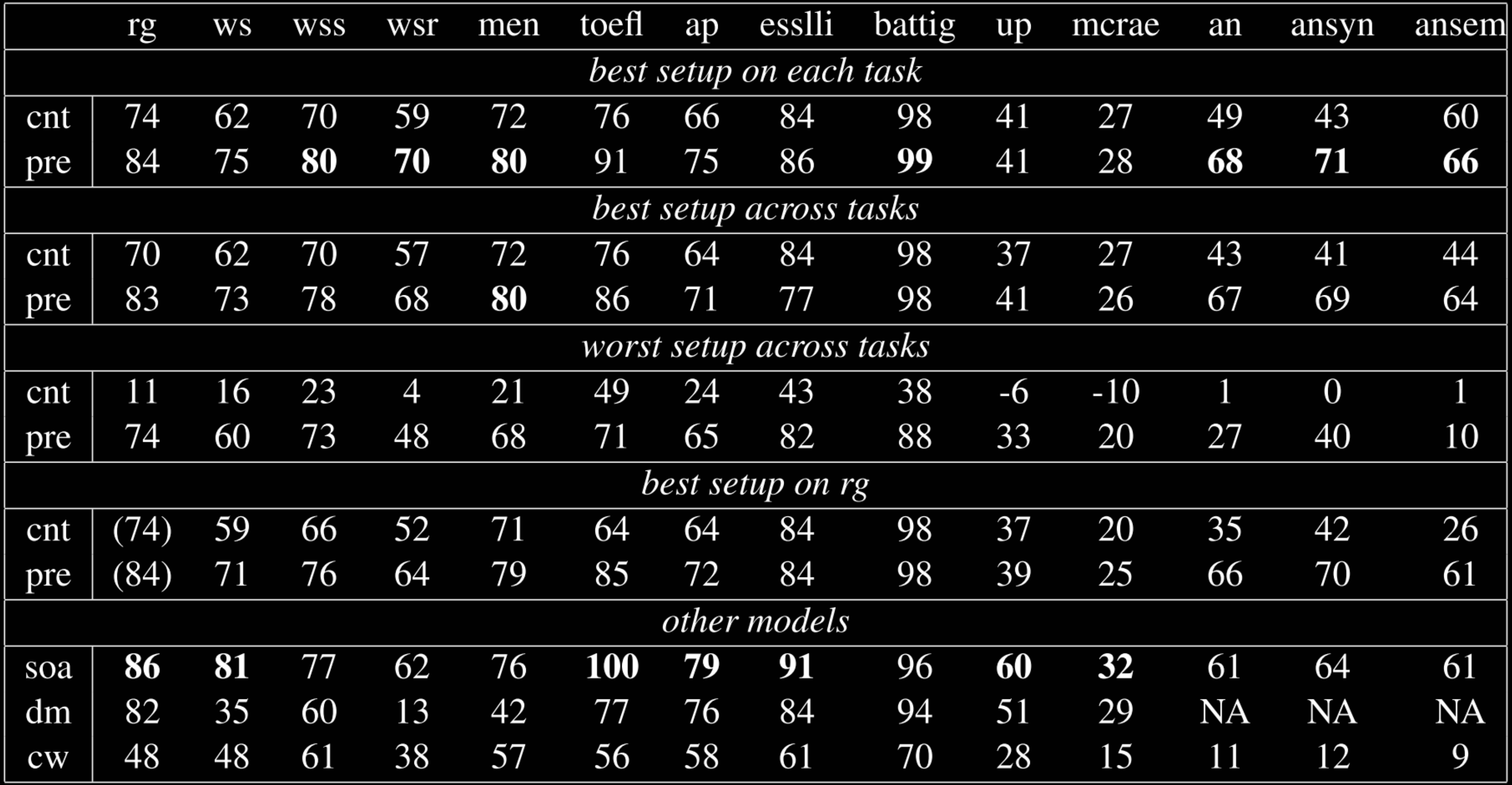
Similarity Evaluation

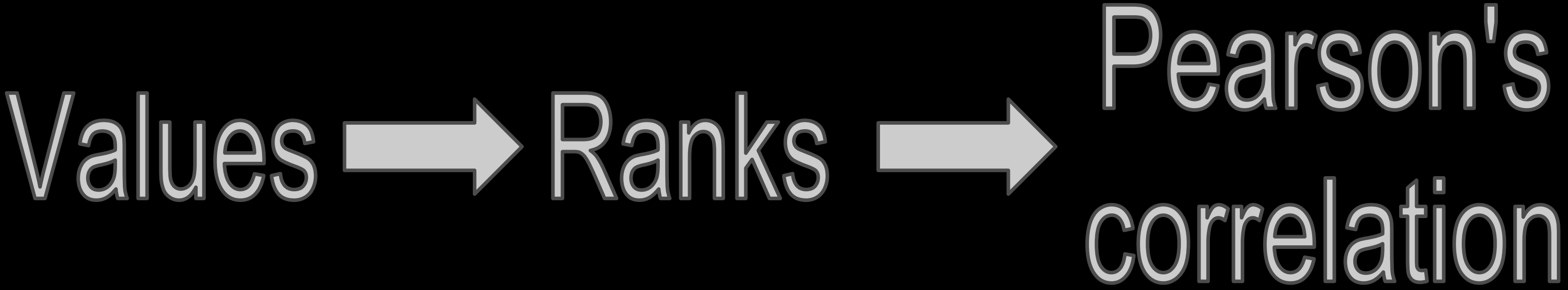
Minh N. Le, Antske Fokkens



	rg	ws	wss	wsr	men	toefl	ap	esslli	battig	up	mcrae	an	ansyn	ansem
best setup on each task														
cnt	74	62	70	59	72	76	66	84	98	41	27	49	43	60
pre	84	75	**80**	**70**	**80**	91	75	86	**99**	41	28	**68**	**71**	**66**
best setup across tasks														
cnt	70	62	70	57	72	76	64	84	98	37	27	43	41	44
pre	83	73	78	68	**80**	86	71	77	98	41	26	67	69	64
worst setup across tasks														
cnt	11	16	23	4	21	49	24	43	38	-6	-10	1	0	1
pre	74	60	73	48	68	71	65	82	88	33	20	27	40	10
best setup on rg														
cnt	(74)	59	66	52	71	64	64	84	98	37	20	35	42	26
pre	(84)	71	76	64	79	85	72	84	98	39	25	66	70	61
other models														
soa	**86**	**81**	77	62	76	**100**	**79**	**91**	96	**60**	**32**	61	64	61
dm	82	35	60	13	42	77	76	84	94	51	29	NA	NA	NA
cw	48	48	61	38	57	56	58	61	70	28	15	11	12	9

From Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don’t count, predict!

Question 1:
What does it mean by ρ=0.6?

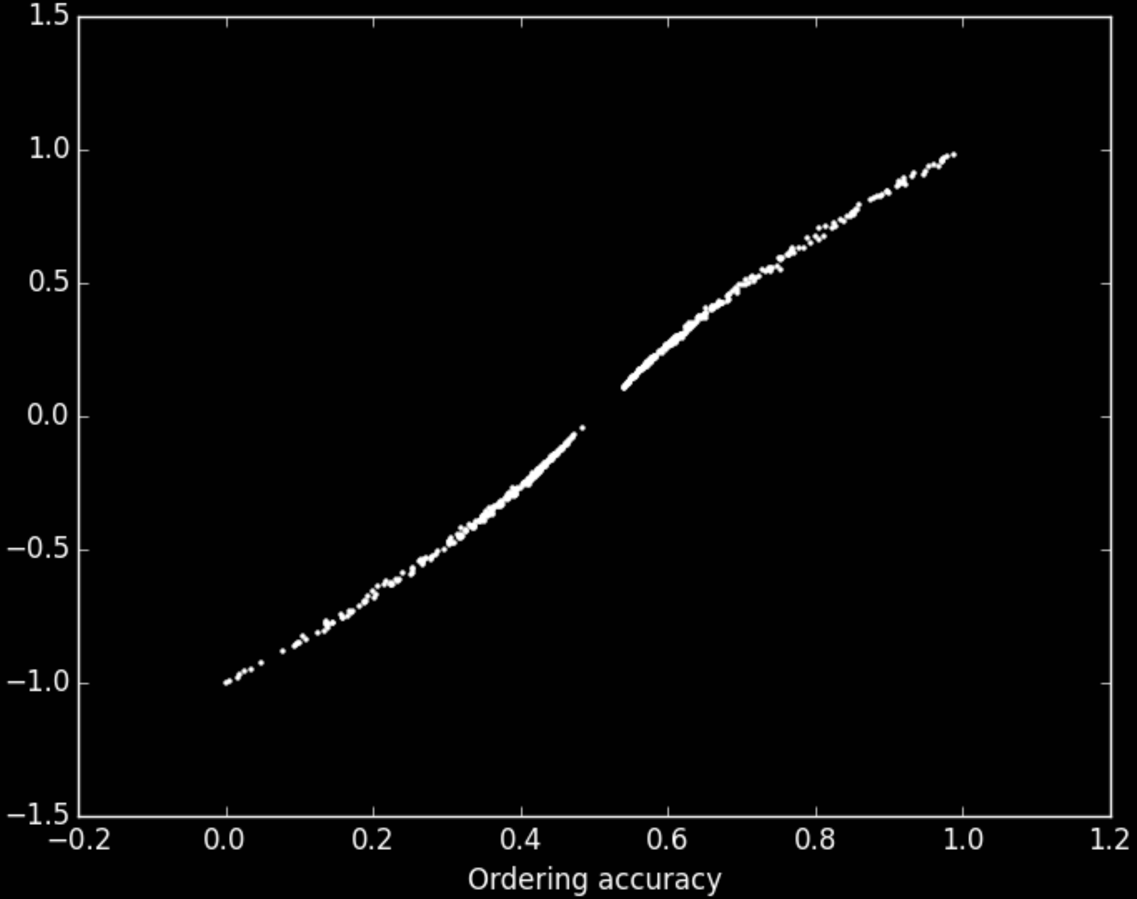


Values → Ranks → Pearson's correlation

Ordering accuracy:

$$a = a_{G,G} = \frac{1}{|G|^2} \sum_{(a,b)\in G} \sum_{(x,y)\in G} 1_{s,G}(a, b, x, y)$$

$$a = \frac{1}{|G|^2} \sum_i \sum_j |g_i||g_j| a_{g_i,g_j}$$


Spearman's rho
Ordering accuracy
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
1.2

Group	Sim.
g_1	0-2
g_2	2-4
g_3	4-6
g_4	6-8
g_5	8-10

Granularity	Pair of groups
0	g_1g_1, g_2g_2, g_3g_3, g_4g_4, g_5g_5
1	g_1g_2, g_2g_3, g_3g_4, g_4g_5
2	g_1g_3, g_2g_4, g_3g_5
3	g_1g_4, g_2g_5
4	g_1g_5

Granularity	Example			Weight
0	take-leave	vs	succeed-try	58%
1	spoon-cup	vs	argue-differ	
2	mad-glad	vs	easy-flexible	42%
3	certain-sure	vs	strong-proud	
4	easy-big	vs	formal-proper	

People can’t reliably judge fine-grained difference in similarity but it is the larger part of Spearman’s rho.

Spearman’s ρ is skewed
towards unreliable comparison
and a big ρ is not necessarily good.

Question 2:
What does it mean by having a similarity of 0.2?

Levels of measurement

- Stevens, S. S. (1946). "On the Theory of Scales of Measurement". Science 103 (2684)
- Details are debatable
- Widely used in papers, books, software

Stevens’ four levels

1. **Nominal:** categories, e.g. noun, verb, adjective, adverb
2. **Ordinal:** rank, e.g. 'completely agree', 'mostly agree', 'mostly disagree', 'completely disagree’
3. **Interval:** degree of difference, e.g. date, Celsius degree
4. **Ratio:** e.g. mass, length, duration,...

Stevens’ four levels

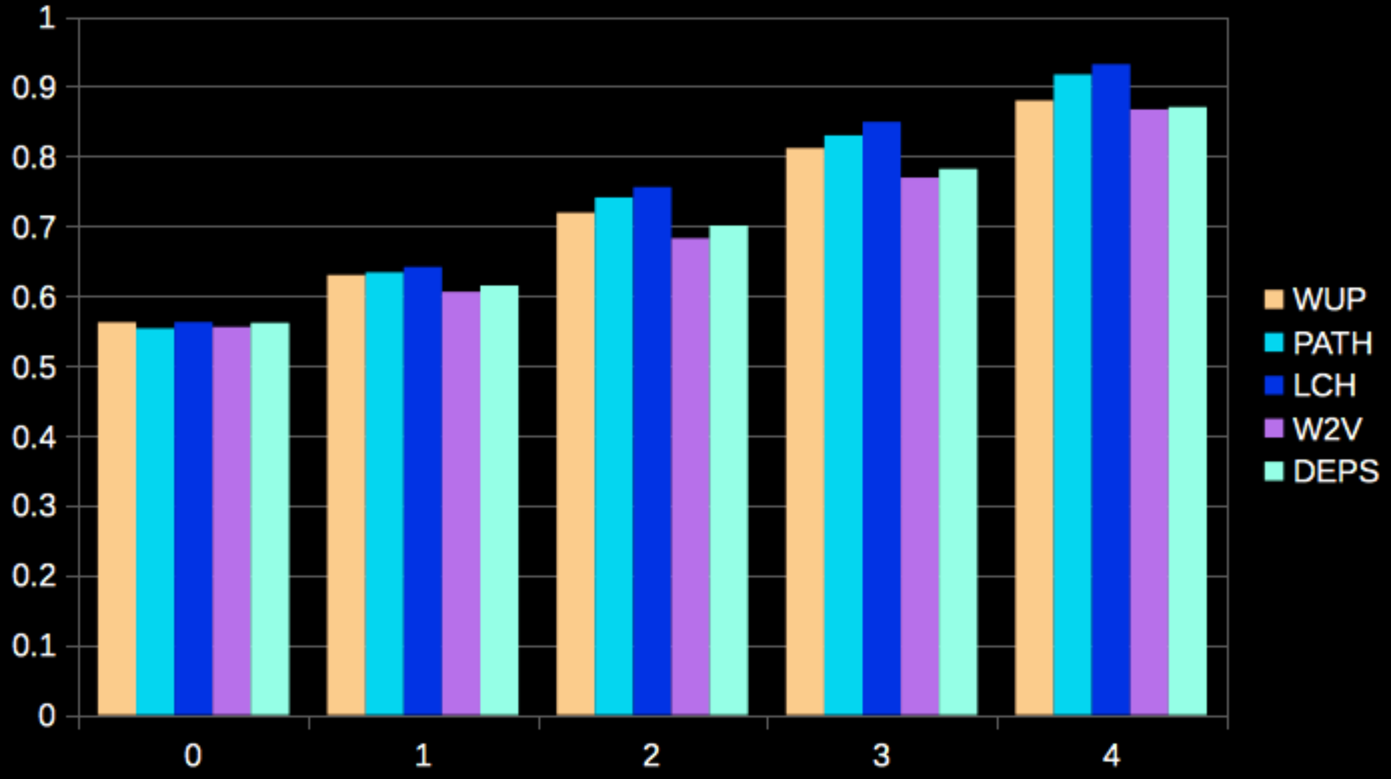
- Later levels allow all mathematical operations of earlier levels but not vice versa
- To compute the mean of some values, they must show “degree of difference”
- We can’t do so with ordinal or nominal values
- We can with interval or ratio values

Is similarity judgment interval/ratio?

- Pairs: P_1: (happy, mad) = 1, P_2: (modest, ashamed) = 2, P_3: (clothes, closet) = 3, P_4: (hand, foot) = 4
- Is the *difference in similarity* between P_1 and P_2 the same as the *difference in similarity* between P_2 and P_3?
- Do P_2 and P_4 *differ twice as much* as P_1 and P_2?

Similarity datasets are based on wrong assumptions and present a distorted view of similarity.

Question 3:
What can we do?


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
WUP
PATH
LCH
W2V
DEPS