Measuring performance: The 2×2 contingency matrix

Black-box or "end-to-end" system performance

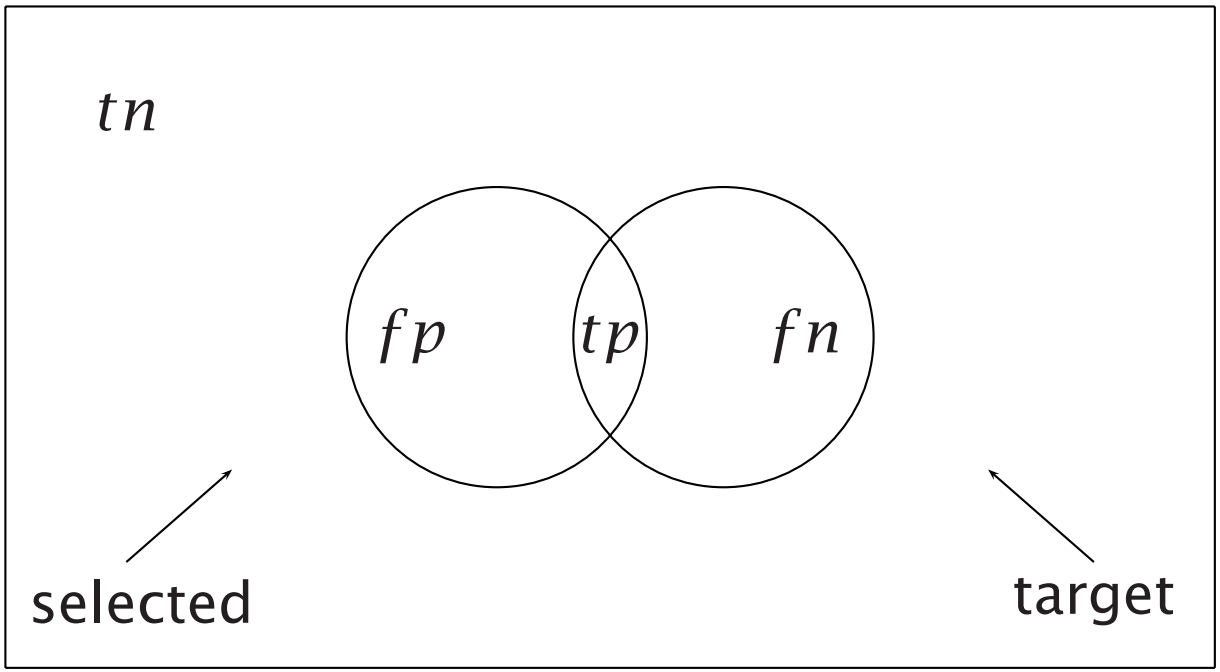
	Actual	
System	target	¬ target
selected	tp	fp
¬selected	fn	tn

Accuracy = $(tp + tn)/N$

Error = $(fn + fp)/N = 1 -$ Accuracy

Why is this measure inadequate for IR?

The motivation for precision and recall



Accuracy is not a useful measure when the target set is a tiny fraction of the total set.

Precision is defined as a measure of the proportion of selected items that the system got right:

$$\text{precision } P = \frac{tp}{tp + fp}$$

Recall is defined as the proportion of the target items that the system selected:

$$\text{recall } R = \frac{tp}{tp + fn}$$

These two measures allow us to distinguish between excluding target items and returning irrelevant items.

They still require human-made “gold standard” judgements.

Evaluation of *ranked* results	Ranking 1	Ranking 2	Ranking 3
	d1: ✓	d10: ×	d6: ×
	d2: ✓	d9: ×	d1: ✓
	d3: ✓	d8: ×	d2: ✓
	d4: ✓	d7: ×	d10: ×
	d5: ✓	d6: ×	d9: ×
	d6: ×	d1: ✓	d3: ✓
	d7: ×	d2: ✓	d5: ✓
	d8: ×	d3: ✓	d4: ✓
	d9: ×	d4: ✓	d7: ×
	d10: ×	d5: ✓	d8: ×
precision at 5	1.0	0.0	0.4
precision at 10	0.5	0.5	0.5
uninterpolated av. prec.	1.0	0.3544	0.5726
interpolated av. prec. (11-point)	1.0	0.5	0.6440

Interpolated average precision